

THE INSIGHT CENTRE FOR DATA ANALYTICS



**XPLODIV: DIVERSIFICATION APPROACH
FOR RECOMMENDER SYSTEMS**

Andrea Barraza-Urbina

Benjamin Heitmann

Conor Hayes

Angela Carrillo Ramos

INSIGHT TECHNICAL REPORT 2015-04-04

JUNE 2014; APRIL 2015

THE INSIGHT CENTRE FOR DATA ANALYTICS

**The Insight Centre for Data
Analytics**

National University of Ireland,
Galway

The DERI Building

IDA Business Park

Lower Dangan

Galway, Ireland

<http://www.insight-centre.org>

XPLODIV: DIVERSIFICATION APPROACH FOR RECOMMENDER SYSTEMS

Andrea Barraza-Urbina¹
Conor Hayes³

Benjamin Heitmann²
Angela Carrillo Ramos⁴

Abstract. Recommender Systems have emerged to guide users in the task of efficiently browsing/exploring a large product space, helping users to quickly identify interesting products. However, suggestions generated with traditional Recommender Systems usually do not produce diverse results, though it has been argued that diversity is a desirable feature. The study of diversity aware Recommender Systems has become an important research challenge in recent years, drawing inspiration from diversification solutions for Information Retrieval. However, we argue it is not enough to adapt Information Retrieval techniques towards Recommender Systems, as they do not place the necessary importance to factors such as serendipity, novelty and discovery which are imperative to Recommender Systems. In this report, we propose a diversification technique for Recommender Systems that generates a diversified list of results which not only balances the trade-off between quality (in terms of accuracy) and diversity, but also considers the trade-off between exploitation of the user profile and exploration of novel products. Our experimental evaluation, composed of both qualitative and quantitative tests, shows that the proposed approach has comparable results to state of the art approaches. Moreover, through control parameters, our approach can be tuned towards more explorative or exploitative recommendations.

Keywords: Diversity, Recommender Systems, Exploitation, Exploration, Information Retrieval, Novelty, Discovery, Relevance

¹ Insight Centre for Data Analytics, National University of Ireland Galway, IDA Business Park, Lower Dangan, Galway, Ireland. E-mail: andrea.barraza@insight-centre.org.

² Insight Centre for Data Analytics, National University of Ireland Galway, IDA Business Park, Lower Dangan, Galway, Ireland. E-mail: benjamin.heitmann@insight-centre.org.

³ College of Engineering and Informatics, National University of Ireland Galway, University Road, Galway, Ireland. E-mail: conor.hayes@insight-centre.org

⁴ Pontificia Universidad Javeriana, Carrera 7 No. 40 - 62 Edificio José Gabriel Maldonado, S.J. Bogotá, Colombia. E-mail: angela.carrillo@javeriana.edu.co.

Acknowledgements: This research was made possible by funding from Science Foundation Ireland under grant number SFI/12/RC/2289 (Insight) and by the Master's Program of the Computer Science Department at the Pontificia Universidad Javeriana, Bogotá.

Submitted as M.Sc. final research project

Copyright © 2015 by the authors

Contents

I. INTRODUCTION	1
1.1 Motivation.....	1
1.2 Research Goal.....	3
1.3 Solution Approach.....	3
1.4 List of Main Contributions.....	4
1.5 Document Structure.....	5
II. BACKGROUND	6
2.1 What is Diversity?.....	6
2.1.1 <i>Cross-field applications of Diversity</i>	6
2.1.2 <i>General properties of Diversity</i>	7
2.1.2.1 Basic Diversity.....	7
2.1.2.2 Dual-concept Diversity.....	8
2.1.2.3 Diversity properties in Stirling’s general framework for analysing diversity [Stir07]..	8
2.1.3 <i>Measuring Diversity</i>	9
2.1.3.1 Basic Diversity.....	10
2.1.3.2 Dual-concept Diversity.....	10
2.1.3.3 Diversity heuristic in Stirling’s general framework for analysing diversity [Stir07]..	11
2.1.4 <i>Dissimilarity metric</i>	13
2.1.5 <i>Diversification Problem</i>	14
2.2 Diversity in Information Retrieval.....	15
2.2.1 <i>Defining Information Retrieval</i>	15
2.2.2 <i>Motivation: Why consider diversity for Information Retrieval?</i>	16
2.2.3 <i>Problem Definition: Diversification Problem in Information Retrieval</i>	17
2.3 Diversity in Recommendation Systems.....	18
2.3.1 <i>Defining Recommendation Systems</i>	18
2.3.1.1 Recommendation Systems vs. Information Retrieval.....	19
2.3.1.2 Types of Recommender Systems.....	20
2.3.2 <i>Motivation: Why consider diversity for Recommender Systems?</i>	21
2.3.2.1 On the uncertainty of the User Profile.....	21
2.3.2.2 Serendipity, Novelty and Diversity.....	22
2.3.3 <i>Definition of research problem: Diversification Problem in Recommendation Systems</i>	22
2.4 Summary.....	25
III. LITERATURE REVIEW	27
3.1 Diversification Techniques.....	27
3.1.1 <i>Diversification Techniques for Information Retrieval</i>	27
3.1.1.1 Maximal Marginal Relevance (MMR) [Carb98].....	27
3.1.1.2 IA-Select algorithm for diversifying search results [Agra09].....	28
3.1.1.3 The xQuAD Framework [Sant10].....	30
3.1.1.4 Coverage-based search result diversification [Zhen12].....	30
3.1.2 <i>Diversification Techniques for Recommendation Systems</i>	32
3.1.2.1 Topic Diversification [Zieg05].....	32

3.1.2.2	Similarity vs. Diversity in Case-based Recommendation Systems [Smyt01].....	33
3.1.2.3	Item Re-Ranking Methods [Adom09]	34
3.1.2.4	Information Retrieval Diversity for Recommender Systems [Varg12]	36
3.1.2.5	Latent Factor Portfolio [Shi12]	37
3.1.2.6	User Profile Partitioning [Zhan09a][Varg13]	38
3.1.2.7	Diversifying User Neighbors [Smyt01][Yang13][Zhan09][Said12].....	39
3.2	Discussion on Diversification Techniques	40
3.3	Evaluation Metrics for Diversity	44
3.3.1	<i>Evaluation metrics for sub-topic retrieval [Zhai03]</i>	44
3.3.2	<i>Intent-Aware Evaluation Measures [Agra09][Varg12]</i>	45
3.3.3	α -nDCG [Clar08] and Novelty and rank-biased precision (NRBP) [Clar09].....	46
3.3.4	<i>Intra-List Similarity [Zieg05]</i>	48
3.3.5	<i>Vargas and Castells formalization of novelty and diversity metrics [Varg11]</i>	48
3.4	Discussion on Evaluation Metrics for Diversity	50
3.5	Summary	52
IV.	EXPLOITATION-EXPLORATION DIVERSIFICATION TECHNIQUE	53
V.	EXPERIMENTAL VALIDATION	57
5.1	Evaluation Environment.....	58
5.2	Qualitative Offline Experiment	60
5.2.1	<i>Discussion on Pure Exploitation Results</i>	60
5.2.2	<i>Discussion on Pure Exploration Results</i>	64
5.3	Quantitative Offline Experiment	68
5.3.1	<i>Experiment Set Up</i>	68
5.3.2	<i>Diversity-Aware Evaluation Framework</i>	69
5.3.3	<i>Data Interpretation and Discussion</i>	73
5.4	Summary	94
VI.	CONCLUSIONS AND FUTURE WORK.....	97
REFERENCES	101

List of Tables

Table 1. Four variants of Stirling's diversity heuristic [Stir07]	13
Table 2. Maximal Marginal Relevance variable definitions.....	28
Table 3. Comparison Criteria for Diversification Techniques.....	41
Table 4. Comparison of Diversification Techniques	44
Table 5. Intent-Aware adaptation of classical IR metrics [Agra09]	46
Table 6. Comparison Criteria for Metrics to Evaluate Diversity.....	50
Table 7. Comparison of Metrics to Evaluate Diversity	51
Table 8. XPLODIV Dimensions.....	54
Table 9. Definition of Implemented Dimensions	59
Table 10. Pure exploitation test combinations.....	60
Table 11. Pure exploitation qualitative test results	64
Table 12. Pure exploration test combinations.....	65
Table 13. Pure exploration qualitative test results	68
Table 14. Average of Results for Test Cases.....	91
Table 15. Answers to Experimental Validation Questions	95

List of Figures

Figure 1. Solution Approach	4
Figure 2. Properties of Diversity according to Stirling [Stir07]	9
Figure 3. Non-parametric measures of ecological diversity. Figure extracted from [Stir98].	11
Figure 4. User-Item Matrix in Recommendation Systems	18
Figure 5. Post-filtering Diversification Techniques	32
Figure 6 Diversity preserving algorithms. Figure extracted from [Smyt01].	34
Figure 7. Re-ranking strategies proposed by [Adom09].....	35
Figure 8. Diversification Techniques	40
Figure 9. XPLODIV Greedy Optimization Algorithm	53
Figure 10. Diversification Techniques Class Diagram	59
Figure 11. Diversity-Aware Evaluation Framework	70
Figure 12. Histogram of Distances to User Profile.....	74
Figure 13. Gini-Simpson Diversity Index — [Graph A] Average Dissimilarity Graph	75
Figure 14. Gini-Simpson Diversity Index — [Graph B] Minimum Dissimilarity Graph.....	76
Figure 15. Gini-Simpson Diversity Index — [Graph C] Tuning Graph	76
Figure 16. Pairwise Intra-List Dissimilarity Metric — [Graph A] Average Dissimilarity Graph	77
Figure 17. Pairwise Intra-List Dissimilarity Metric — [Graph B] Minimum Dissimilarity Graph.....	78
Figure 18. Pairwise Intra-List Dissimilarity Metric — [Graph C] Tuning Graph	78
Figure 19. Normalized Discounted Cumulative Gain — [Graph A] Average Dissimilarity Graph.....	79
Figure 20. Normalized Discounted Cumulative Gain — [Graph B] Minimum Dissimilarity Graph.....	80
Figure 21. Normalized Discounted Cumulative Gain — [Graph C] Tuning Graph	80
Figure 22. User Profile Exploitation Metric — [Graph A] Average Dissimilarity Graph.....	81
Figure 23. User Profile Exploitation Metric — [Graph B] Minimum Dissimilarity Graph	82
Figure 24. User Profile Exploitation Metric — [Graph C] Tuning Graph.....	82
Figure 25. Average User Profile Similarity Metric — [Graph A] Average Dissimilarity Graph	83
Figure 26. Average User Profile Similarity Metric — [Graph B] Minimum Dissimilarity Graph.....	84
Figure 27. Average User Profile Similarity Metric — [Graph C] Tuning Graph	84
Figure 28. Dissimilarity Threshold Percentage Metric — [Graph A] Average Dissimilarity Graph	85
Figure 29. Dissimilarity Threshold Percentage Metric—[Graph B]Minimum Dissimilarity Graph	85
Figure 30. Dissimilarity Threshold Percentage Metric — [Graph C] Tuning Graph	86
Figure 31. Number of Categories in List Metric — [Graph A] Average Dissimilarity Graph	87
Figure 32. Number of Categories in List Metric — [Graph B] Minimum Dissimilarity Graph.....	87
Figure 33. Number of Categories in List Metric — [Graph C] Tuning Graph.....	88
Figure 34. Percentage of Item Replacements Metric — [Graph A] Average Dissimilarity Graph	89
Figure 35. Percentage of Item Replacements Metric — [Graph B] Minimum Dissimilarity Graph	89
Figure 36. Percentage of Item Replacements Metric — [Graph C] Tuning Graph	90
Figure 37. Integrated view of Evaluation Perspectives	92
Figure 38. Win-Loss comparison of Evaluation Perspectives	92

List of Equations

Equation 1. Diversity measured as an aggregate of pairwise dissimilarities	10
Equation 2. Diversity measured as the average of pairwise dissimilarities	10
Equation 3. Diversity that an element p would add to a set \mathcal{S}	10
Equation 4. Stirling's diversity heuristic [Stir07]	12
Equation 5. Transformations of similarity metric into distance metric [Chen09]	14
Equation 6. Distance as the inverse of similarity.....	14
Equation 7. Nemhauser bound for greedy optimization of a submodular function [Nemh78].....	18
Equation 8. Greedy Algorithm Optimization Function [Zhen12].....	27
Equation 9. Maximal Marginal Relevance	28
Equation 10. Agrawal diversification optimization function [Agra09]	29
Equation 11. Marginal gain in IA-Select algorithm [Agra09].....	29
Equation 12. Update of $U(c q, \mathcal{S})$ in IA-Select algorithm [Agra09]	29
Equation 13. xQuAD framework objective function [Sant10]	30
Equation 14. xQuAD diversity probability [Sant10]	30
Equation 15. xQuAD framework objective function rewritten [Sant10].....	30
Equation 16. Coverage-based framework for diversification in IR [Zhen12]	31
Equation 17. Squared loss coverage function [Zhen12]	31
Equation 18. Defining coverage and weight functions [Zhen12].....	31
Equation 19. Topic Diversification [Zieg05].....	33
Equation 20. Dissimilarity rank for Topic Diversification [Zieg05]	33
Equation 21. Quality metrics in [Smyt01]	34
Equation 22. Re-ranking Method proposed by [Adom09]	35
Equation 23. α_i in Re-ranking Method proposed by [Adom09]	35
Equation 24. Modified Marginal gain from IA-Select [Agra09][Varg12]	36
Equation 25. Adaptation of IA-Select using the aspect space [Agra09][Varg12]	36
Equation 26. Modified Marginal gain from IA-Select [Agra09][Varg12]	37
Equation 27. Sub-topic recall at rank K [Zhai03].....	45
Equation 28. Intent-aware nDCG adapted to RecSys using the concept of aspect space [Varg12]	45
Equation 29. Probability a document is relevant to a user [Clar08]	46
Equation 30. Probability a document contains a nugget [Clar08]	47
Equation 31. Probability a document is relevant to a user [Clar08]	47
Equation 32. Probability a document at rank k is relevant to a user [Clar08]	47
Equation 33. Gain of a document at rank k for a user [Clar08].....	47
Equation 34. Discounted Cumulative Gain at rank k [Clar08].....	47
Equation 35. α - Normalized Discounted Cumulative Gain at rank k [Clar08]	48
Equation 36. Novelty- and rank-biased precision (NRBP) [Clar09]	48
Equation 37. Re-ranking Method proposed by [Adom09]	48
Equation 38. General specification of Vargas framework for diversity metrics [Varg11]	49
Equation 39. Rank-sensitive and relevance-aware diversity metric [Varg11].....	49
Equation 40. Exploitation-Exploration diversification XPLODIV	53
Equation 41. Relevance Dimension.....	54
Equation 42. Example of Diversity that an element i would add to a set \mathbb{R}	54

Equation 43. Average pairwise distance of an element i to a set \mathbb{R}	55
Equation 44. Exploitation dimension as probability of high rating of similar items	55
Equation 45. Exploitation dimension as probability of high rating of nearest neighbors	55
Equation 46. Exploration as diversity of item i to the user profile \mathbb{U}	56
Equation 47. Exploration as diversity of item i to the nearest neighbors	56
Equation 48. User Profile Heterogeneity for MovieLens dataset	59
Equation 49. Normalized Discounted Cumulative Gain (nDCG)	71
Equation 50. User Profile Exploitation metric.....	72
Equation 51. Dissimilarity Threshold Percentage metric	72

Chapter I

INTRODUCTION

In this work, we address the diversification problem in Recommendation Systems by proposing a novel diversification technique that balances aspects related to relevance, diversity, exploitation of the user profile and exploration of novel products. In the first place, we present an overview of the motivation of our work. Next, we formalize the research goal of the project. Subsequently, we describe the solution approach carried out in order to achieve the research goal. Following, we describe the main contributions of our work, which are: (i) a novel *Exploitation-Exploration diversification approach* called *XPLODIV*, (ii) a Diversity-Aware Evaluation Framework that organizes and defines metrics to evaluate Recommendation Systems within the context of diversity, and (iii) an analytical comparison of Related Work which identifies advantages and disadvantages of current approaches. Lastly, we describe the structure of this document.

1.1 Motivation

Diversity is a concept that has been applied in many fields; mostly with the goal of obtaining a set of objects that have a high level of *dissimilarity* between them, and that as a group, maximize a quality criterion. However, there is usually a trade-off between diversity and quality; hence, the *diversification problem* is how to choose k elements from a set that maximizes diversity at a low quality sacrifice.

Extensive work on the field of Information Retrieval (*IR*), has been carried out to solidify concepts related to the diversification problem. The study of diversity as it has been applied in *IR* serves as a strong foundation for work on diversity in Recommendation Systems (*RecSys*). Thus, the aim of this section is to introduce the motivation surrounding research on diversity for both *IR* and *RecSys*.

In *IR*, diversity is a highly desirable feature. In the first place, diversity aims at removing redundancy within the retrieved results. Redundancy is found because in *IR* the document search space usually contains highly duplicative information, and thus documents that are similar to the target query also tend to be similar to each other [Carb98][Ceri13]. As a consequence, *IR* systems without diversification usually provide users an over-specialized homogenous set of results. This is not desirable, it would mean that if one document is not relevant then all similar results are not relevant, and in consequence, there is a high risk of not satisfying the user [Ceri13]. In second place, diversity is used as a response to query ambiguity. Without further information to help determine the precise user intent, *IR* results should have high coverage of the different interpretations of the query to increase the chance of satisfying a user with a random intent [Ceri13][Zhen12].

In this fashion, the goal of diversification in *IR* is to select documents that are not only relevant to the query but that also cover as many query interpretations as possible. However, there is a trade-off between selecting items that are of higher relevance (which tend to be similar to each other) and obtaining diverse results [Goll09]. Therefore, the diversification problem in *IR* is

usually modelled as a bi-criteria optimization problem that aims to find the appropriate balance between two competing objectives: maximizing diversity and maximizing relevance [Goll09].

In *RecSys*, diversity is also a highly desirable feature.

On the one hand, diversity is important to deal with the uncertainty surrounding the user profile. The only evidence of user tastes/likes a *RecSys* has is encapsulated within the user profile. However, much like a user query in *IR*, the user profile could be incomplete and ambiguous. This can be explained by: the large size of item spaces and the unfeasibility of obtaining explicit rating information on all products from users, the unreliability of interpreting implicit information to understand user likes, and the dynamic nature of user preferences. In face of user profile uncertainty, *RecSys* should offer users a diverse set of suggestions representative of the variety of the user's tastes in order to increase the chances the user finds useful items in recommendations [Varg12][Zhen12].

On the other hand, diversity is essential to the concept of novelty, which is directly related to the idea of discovery and essential to the purpose of *RecSys*. The relation between novelty and diversity, is established on the notion that different levels of novelty can be achieved depending on how far or *diverse* an item is from the user's past experience. In addition to aiding discovery, novel recommendations help increase the information flow between the user and the system. It is to be expected that discovering new products would lead to an information gain for the user, but this is also true for the *RecSys* itself. Discovery of new items leads to user feedback on diverse/novel items. This feedback generates larger information gain for the user profile than feedback of non-novel items, broadening the knowledge over the user preferences [Lemi08].

Even though diversity is a desirable feature, *RecSys* do not offer diverse recommendations naturally. This is due to: (a) *the heuristics that lay foundation to RecSys techniques are based on similarity measures*: traditional techniques that are centered on similarity-based heuristics suffer problems like overspecialization, bias towards popular items, and bias towards items which are similar to highly-rated items from the user profile; (b) *traditional evaluation metrics encourage accuracy but penalize diversity*: with traditional *RecSys* techniques novel products tend to receive lower predicted ratings compared to products similar to those the user always consumes, as a consequence, accuracy metrics penalize recommending novel products; and (c) *recommendation list evaluation is performed as an aggregate of the individual scores of items, disregarding the real value of items in the context of the list*: recommendation list metrics do not evaluate each product within the context of the list and cannot determine if the list offers items that are both of high quality and sufficiently diverse to cover the spectrum of the user's interests [Mcne06].

It can be seen that the diversification problem in *RecSys* is similar to that in *IR*, where there is a trade-off between the individual accuracy of an item and the overall diversity of the recommendation list. In this manner, the diversification goal in *RecSys* would be to generate a list of suggested items that maximize both the predicted rating for items and coverage over the wide spectrum of user preferences. However, *RecSys* must also account for novel products, which by definition are not directly related to the identified user preferences. This brings up an additional trade-off between how much the *RecSys* wants to *exploit* the known information about the user by covering the preferences in the user profile, and how much the *RecSys* wants to *explore* what other preferences the user could have by offering novel products.

The trade-off between exploitation and exploration could depend on many factors, such as the maturity of the user profile and the user's openness to experience. For example, for a new user,

the *RecSys* might want to offer more novel/explorative products in order to gain information about the user's interests. In contrast, for users that are not very open to new experiences, they may possibly prefer to receive recommendations of products similar to those they have liked in the past (*i.e.*, exploitative items). As a result, it is important for the diversification technique in *RecSys* to be tunable, so in this way it can be adapted to the *RecSys* requirements of diversity, exploitation and exploration.

In this section, we present a summary of the motivation surrounding the diversification problem as it has been defined for *IR* and *RecSys*. In the following section, we will formalize the research goal of this work.

1.2 Research Goal

The main research goal that we will address in this work is:

Design a diversification technique for *RecSys* that can balance the trade-off between quality (in terms of relevance) and diversity, considering the trade-off between *exploitation* of the user profile and *exploration* of novel products.

In the following section, we will discuss the solution approach carried out towards achieving the research goal.

1.3 Solution Approach

In order to address the research goal, in this work, we propose a novel diversification technique. In the first place, we carried out a comprehensive literature review to identify the advantages and disadvantages of current diversification solutions. From the literature review, we extracted the foundations for the diversification technique to be proposed. Next, taking into account desirable criteria for a diversification solution identified from the literature review, we designed the proposed Exploitation-Exploration diversification approach. In order to verify that the approach fulfills the promised features, we carried out experimental validation of the proposed diversification technique. As a first step for experimental validation, we carried out another literature review, this time to identify useful metrics to evaluate our work. We found that current evaluation metrics were insufficient for our purposes and thus, we proposed a novel Diversity-Aware Evaluation Framework that would allow us to verify the functionality of our diversification technique. Finally, we ran two types of tests: qualitative tests and quantitative tests—which were carried out using the evaluation methodology proposed in the Diversity-Aware Evaluation Framework—. From evaluation results, we provided evidence that showed that our approach can be tuned using the control parameters and that it also generates results comparable to baselines and state-of-the-art techniques.

In brief, our solution approach is composed of the following four steps (view *Figure 1*):

- (i) *Literature Review*: survey of current diversification techniques and diversification evaluation strategies.
- (ii) *Design*: design of the proposed diversification technique.
- (iii) *Development*: implementation of a functional prototype of our diversification approach, baselines and state-of-the-art approaches
- (iv) *Experimental Validation*: analysis of results obtained from both quantitative and qualitative tests.

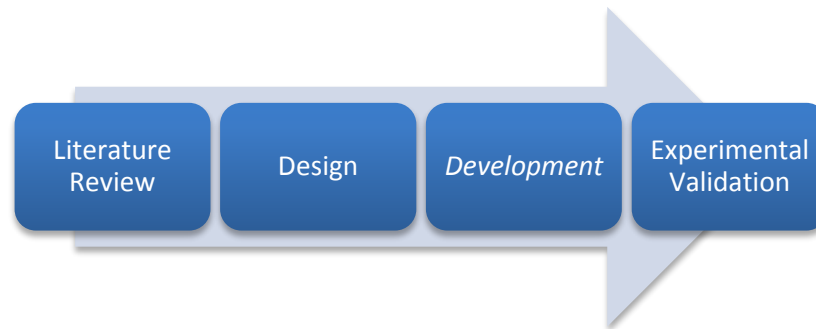


Figure 1. Solution Approach

In this section, we have presented the solution approach of this work. In the following section, we will highlight the main contributions of our project.

1.4 List of Main Contributions

In our work we highlight the following main contributions:

Exploitation-Exploration diversification approach (XPLODIV)

We propose a novel diversification technique called *XPLODIV*, which considers not only the trade-off between relevance *vs.* diversity, but also the trade-off between exploitation *vs.* exploration. Through experimental validation, we show that our approach can be tuned using the provided control parameters, and that it produces results comparable to baselines and state-of-the-art techniques.

Overall, our approach presents an improvement over current solutions, as it can be tuned towards more diverse exploitative results or more diverse explorative results. A crucial differentiating factor, is that *XPLODIV* explicitly accounts for the amount of novelty that is imprinted on a recommendation list. In this way, we address control over indispensable aspects of recommender results related to discovery and serendipity, which are disregarded in current approaches. Altogether, with *XPLODIV*, a Recommendation System application can be adjusted towards the requirements of the use context and user preferences on diversity, exploitation and exploration.

Diversity-Aware Evaluation Framework for Recommender Systems

We specify a Diversity-Aware Evaluation Framework, which identifies and organizes metrics that should be taken into account when evaluating Recommendation Systems within the context of diversity. After a study of related work, we found the need for a framework to structure the individual existing metrics, and to allow for the incorporation of new metrics, under the idea of evaluating results from different separate perspectives. We believe that proposing an integrated metric, which would offer a joint evaluation of all recommendation aspects at the same time — aspects such as relevance, diversity, exploitation and exploration—, is not only extremely complex to design, but would not allow us to observe the distinct characteristics of aspects and compare these, in order to better understand the possible trade-offs between pairs of aspects. Our framework, is an alternative to an integrated metric, which establishes different evaluation perspectives that individually provide a view on the value of an aspect related to a recommendation list. Perspectives should be first analyzed individually, and then in comparison to each other, and in this fashion obtain an integrated interpretation of the quality of results. We

associated each perspective to a number of evaluation metrics, some of these referenced from current work and others proposed by us in response to the evaluation needs of the work.

Analytical comparison of Related Work

We carry out a robust analytical comparison of related work, which to our knowledge, has not been carried out before. In this analysis, we considered works from both Information Retrieval and Recommendation Systems. We separately examined studies that proposed diversification techniques and studies that proposed diversity-related evaluation metrics. In our comparative analysis, we defined desirable criteria that a proposed solution should have, and evaluated current works on the established criteria. By comparing works from both the fields of Information Retrieval and Recommendation Systems, we were able to analyze the relation between the two fields of research within the context of the diversification problem. Findings from our literature review improve knowledge on the field of diversification for Recommender Systems, by emphasizing the advantages and disadvantages of current approaches, and serve as valuable reference for future work.

In this section, we have presented the main contributions of our work. In the following section, the document structure is described.

1.5 Document Structure

The remainder of this document is organized as follows:

Chapter II. This chapter, provides the conceptual foundations of this research. Particularly, we define the diversification problem for both Information Retrieval and Recommendation Systems. We study Information Retrieval diversification techniques given that they serve as inspiration for Recommendation Systems techniques.

Chapter III. In this chapter, a literature review of both diversification techniques and diversity evaluation metrics is offered for both Information Retrieval and Recommendation Systems. In this review, we offer comparative analysis of related work, emphasizing advantages and disadvantages of current approaches.

Chapter IV. This chapter, presents the Exploitation-Exploration diversification approach named *XPLDIV*. This technique, is composed of four core dimensions, for which we present individual descriptions and implementation details.

Chapter V. In this chapter, we analyze results from the experimental validation to show the effectiveness and tunability of our approach. Specifically, we analyze quantitative and qualitative tests. Furthermore, we propose the Diversity-Aware Evaluation Framework as the methodological ground of quantitative tests.

Chapter VI. Lastly, in this chapter, we conclude and highlight future work.

Chapter II

BACKGROUND

This chapter provides a concise literature review of the basic concepts and definitions that laid foundation to this research. Concretely, an analysis of Diversity and its meaning within both Information Retrieval and Recommendation Systems is covered. We analyze these two fields because the study of diversity as it has been applied in Information Retrieval serves as a strong foundation for work on diversity in Recommender Systems. Finally, a synthesis that connects all the reviewed topics is presented. It should be made clear that each of the discussed areas has a vast body of literature and can be subject to separate projects in themselves. This chapter only covers a brief introduction to each concept.

Specifically, the chapter is organized as follows: (i) *Section 2.1* offers a cross-field generic analysis of the concept of Diversity, (ii) *Section 2.2* explores the application of Diversity to the field of Information Retrieval (iii) *Section 2.3* explores Diversity within the domain of Recommender Systems, lastly (iv) *Section 2.4* offers a synthesis of the discussed concepts.

2.1 What is Diversity?

Diversity is a concept that has been applied in many fields; mostly with the goal of obtaining a set of objects that have a high level of *dissimilarity* between them, and that as a group, maximize a quality criterion. Despite its many applications there has been little work that attempts to characterize the overall concept of diversity in a way that is generically applicable [Stir07]. Even though a profound study of the conceptualization of diversity is not within the scope of this project, it is important to lay out the foundations of the concept and establish desirable properties for a diversity-conscious application.

In this section, we describe works that have proposed a generic representation of diversity. In each section, we address the following questions: (i) *Section 2.1.1*: How is diversity currently applied?, (ii) *Section 2.1.2*: What are the different properties/dimensions of diversity?, (iii) *Section 2.1.3*: How can diversity be measured?, (iv) *Section 2.1.4*: What are the properties of a dissimilarity metric?, and (v) *Section 2.1.5*: Why adding diversity is a challenge?.

The following section offers a simple overview of the application of diversity in various fields.

2.1.1 Cross-field applications of Diversity

The concept of diversity has been studied in various fields, some of which are ecology, economy, urban planning and information sciences [Stir07]:

“Indeed, the generality of the concept almost guarantees its wide application.” [Mcdo03]

Some applications are:

- *Biodiversity*: Nehring *et al.* [Nehr02] define that *“the value of diversity consists in the realization of certain attributes/potentialities of life by some existing species; examples of such attributes are “being a primate”, “a carnivore”*”. Conservation of a diverse and representative sample of species is desirable because it is uncertain which species, that may

not be valuable now, could prove valuable in the future as knowledge grows [Mac108]. This can be referred to as the “*option value*” of biodiversity [Mac108].

- *Statistical surveying*: a sample of the population needs to be chosen that covers/represents the various attributes observed in the population as a whole [Lemi08]. In this way, a sample would represent a “*summary description of a population with a class structure*” [Jung94].
- *Modern Portfolio Theory*: adding diversity is an effective form of reducing the risk of a portfolio by: “*not putting all of your eggs in one basket*” [Wang09]. Given the uncertainty of financial investments and if asset values are not perfectly positively correlated, a diverse portfolio offers more flexibility thus reducing risk [Stir07].
- *Facility Dispersion*: facility dispersion problems, from the area of Operations Research, deal with locating “*obnoxious facilities*” as far away from each other as possible. These “*obnoxious facilities*” are places like nuclear-power plants, oil-storage tanks, and ammunition dumps which should be far from each other in case of accident [Zucc12].

In general, diversity tends to be a desirable feature in domains where decisions need to be made in face of uncertainty or lack of information [Stir98].

2.1.2 General properties of Diversity

In this section, an overview of works that have defined the properties for diversity is offered. These works are divided in three sections according to the complexity of the definition for properties.

2.1.2.1 Basic Diversity

A basic definition for diversity is: *A set is diverse if there is a high level of heterogeneity (dissimilarity) between the items in the collection.*

With this simple definition, there are various ways to measure the diversity of a set; some are defined in *section 2.1.3.1*. In this case, the main property of diversity is *dissimilarity*. In other words, set diversity is defined as the representation of how items are different among themselves.

The most common way to measure diversity, founded on this basic definition, is to quantify it as an aggregate of the pairwise dissimilarity between the items in the set [Nehr02][Lemi08][Hurl11].

As an alternative, Nehring *et al.* [Nehr02] develop basic intuitions for a theory of diversity and define a multi-attribute approach to measure diversity: “*the diversity of a set is simply taken to be the sum of the numeric values (“weights”) of the attributes realized by some object in the set.*” However, applying this approach can be a challenge if there is not a clear set of attributes or a defined taxonomy for objects.

A different approach is to measure diversity as the average *rarity* of the elements in the set [Pati82]. For Patil *et al.* [Pati82] an item is rare if it can rarely be found within the set. They place an example where a traveler wants to find an object in the set that is similar to another he has seen before. The traveler goes through each item one by one; if the set is diverse it will be difficult for the traveler to find a similar object.

Indeed, basic diversity is a desirable characteristic when a representative sample of objects from a broader collection is needed. The most common way of measuring dissimilarity between two items is as the inverse of their similarity [Lemi08]. If we measure similarity as the degree

to which two objects share the same features/attributes, then dissimilarity would measure the degree to which two objects have unique features compared to each other. Therefore, a diverse collection would contain items that have a high proportion of unique features and thus higher coverage of overall features [Lemi08].

2.1.2.2 *Dual-concept Diversity*

Junge [Jung94] defines a dual-concept of diversity that considers two dimensions: (i) the number of discrete categories that elements in a set can be classified into, and (ii) the evenness (homogeneity) of the apportionment of the elements across categories. The second dimension is usually overlooked but Junge argues that a set is most diverse when it reflects both dimensions. In other words, when it considers both objects from many categories and the distribution of elements among categories is flat or even, indicating that all categories are equally represented. In this manner, “*diversity becomes an interaction of the number of categories with the assignment of elements to those categories*” [Mcdo03].

2.1.2.3 *Diversity properties in Stirling’s general framework for analysing diversity* [Stir07]

Stirling on [Stir07] proposes a general framework for analysing diversity that can serve as a reference applicable across a variety of fields. He explains that even though diversity has been worked on different areas of science, a general characterization of diversity has yet to be defined. To this end, he studies diversity in a cross-disciplinary way in order to identify similar underlying properties for diversity from its existing applications on different fields.

Stirling determines that diversity is *an attribute of any system whose elements can be assigned to different categories*. In this manner, diversity concepts have a combination of three basic properties:

- (a) *Variety*: refers to “*the number of categories into which system elements are apportioned. It is the answer to the question: ‘how many types of thing do we have?’*” [Stir07].
- (b) *Balance*: is “*a function of the pattern of apportionment of elements across categories. It is the answer to the question: ‘how much of each type of thing do we have?’*” [Stir07]. The more even the distribution, the more balance.
- (c) *Disparity*: refers to “*the manner and degree in which the elements may be distinguished. It is the answer to the question: ‘how different from each other are the types of thing that we have?’*”. Disparity is usually based on a form of distance measure, and its characterization implicitly determines variety and balance.

For example, in the context of marketing each property would answer the following:

- (a) *Variety*: in how many categories can available products be classified into?
- (b) *Balance*: how many products exist for each category? Another perspective would be to ask: do available products have the same market share?
- (c) *Disparity*: are available products different from one another? How different are categories from each other?

The relations between the properties to diversity can be viewed in *Figure 2*, and are clarified as follows [Stir07]:

- “*All else being equal, the greater the variety, the greater the diversity.*”
- “*All else being equal, the more even is the balance, the greater the diversity.*”

- “All else being equal, the more disparate are the represented elements, the greater the diversity.”

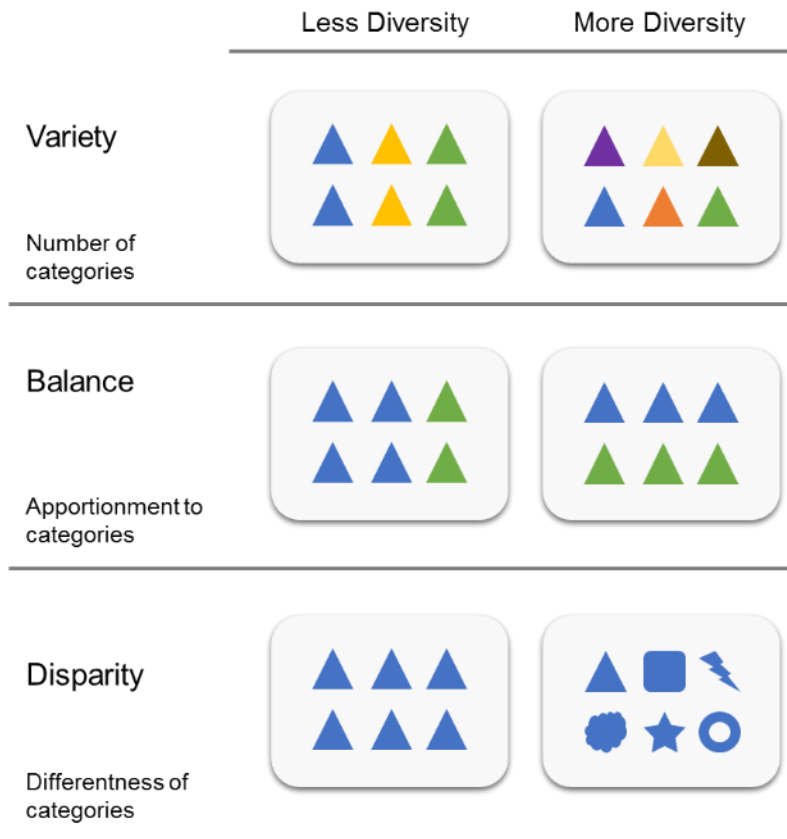


Figure 2. Properties of Diversity according to Stirling [Stir07]

Stirling argues that each property is necessary to define diversity but insufficient on its own. For example, a set can be disparate by containing many types of objects, but unbalanced if objects of one of these types represent a considerably higher portion of the set. In this case the set has high disparity but low balance.

2.1.3 Measuring Diversity

In this section, an overview of metrics used to measure diversity is presented.

To begin, some general rules for the defined metrics are clarified. Lemire *et al.* [Lemi08] define the following rules or axioms for diversity measures in general:

- The diversity of a set containing one element is zero.
- The measure of diversity is a non-negative number.
- Diversity is monotonic, which means that adding a new object will not decrease the diversity of the set.

Nehring *et al.* [Nehr02] add:

- Diversity is a *submodular* function¹. This means, as the set increases it will be harder for an object to add to the diversity of the set.

¹ Refer to [Nemh78] for a definition of submodular function.

This section will discuss metrics at two levels: (i) considering the basic properties discussed in *section 2.1.2.1* and, (ii) considering Stirling's properties discussed in *section 2.1.2.3*.

2.1.3.1 Basic Diversity

In this section diversity measures that only consider the property of *dissimilarity* between the elements are overviewed.

As was mentioned in *section 2.1.2.1*, diversity can be measured as an aggregate of the pairwise dissimilarity between the elements of a set. For a set of size N , $\mathbb{S} = \{s_1, s_2, \dots, s_N\}$ the diversity of the set can be written as in *Equation 1*.

$$\text{diversity}(\mathbb{S}) = \underset{\substack{s_i \in \mathbb{S}, \\ s_j \in \mathbb{S}, i \neq j}}{\text{aggregate}} d(s_i, s_j) \quad (1)$$

Equation 1. Diversity measured as an aggregate of pairwise dissimilarities

Where $d(s_i, s_j)$ is a function that measures the dissimilarity between the objects s_i and s_j . Dissimilarity metrics are discussed in *section 2.1.4*.

Several works such as [McSh02],[Smyt01], and [Zhan08] propose the average function as their aggregation function as in *Equation 2*.

$$\text{diversity}(\mathbb{S}) = \frac{2}{N(N-1)} \left(\sum_{\substack{i=1, \\ s_i \in \mathbb{S}}}^{N-1} \sum_{\substack{j=i+1, j \neq i, \\ s_j \in \mathbb{S}}}^N d(s_i, s_j) \right) \quad (2)$$

Equation 2. Diversity measured as the average of pairwise dissimilarities

When attempting to define diversity, Weitzman [Weit92] first defines the amount of diversity that would be gained for a set if a new element is added (view *Equation 3a*). He further defines that the amount of diversity that an element p would add to the set \mathbb{S} is defined as the distance from point p to the set \mathbb{S} [Weit92][Nehr02]. In this way, Weitzman [Weit92] states that the amount of diversity that an element p would add to a set \mathbb{S} is equal to the dissimilarity between p and it's "nearest neighbor" in \mathbb{S} , as in *Equation 3b*. Consequently, if p is closely related to an element in \mathbb{S} then little diversity is added, and if p is very different from all elements in \mathbb{S} then a lot of diversity is added.

$$\text{distance}(p, \mathbb{S}) = \text{diversity}(\mathbb{S} \cup \{p\}) - \text{diversity}(\mathbb{S}) \quad (3a)$$

$$\text{distance}(p, \mathbb{S}) = \min_{s_i \in \mathbb{S}} d(s_i, p) \quad (3b)$$

Equation 3. Diversity that an element p would add to a set \mathbb{S}

2.1.3.2 Dual-concept Diversity

Stirling in [Stir98] examines a number of diversity metrics and discovers that none account for the three diversity properties he has proposed (*i.e.*, variety, balance and disparity) at the same time. In *Figure 3*, a number of diversity metrics are presented, classified by the properties they consider. It can be seen that none of the metrics consider the property of disparity.

Notation	Ecological Meanings	Interpretation for Economic Systems
A	defined area	defined system parameter
ln	logarithm (usually natural)	logarithm (usually natural)
N	total number of individuals	total system scale
N_i	number of individuals of species i	scale of option i
N_{max}	number of individuals in most populous species	scale of dominant option in portfolio
n	number of individuals in sample	<i>sampling unlikely to be employed</i>
n_i	number of individuals of species i in sample	<i>sampling unlikely to be employed</i>
p_i	proportion of all individuals in species i	proportional contribution of option i
S	number of species	number of options

Index of	Index Name (and source reference)	Diversity =
variety	Species Count (eg: MacArthur, 1965)	S
	Numerical Richness (eg: Magurran, 1988)	$\frac{S}{N}$
	Numerical Richness (eg: Odum et al, 1960)	$\frac{S}{\ln N}$
	Margalef (1958)	$\frac{S - 1}{\ln N}$
	Menhinnick (1964)	$\frac{S}{\sqrt{N}}$
	Species Density (eg: Magurran, 1988)	$\frac{S}{A}$
	Species Density (Gleason, 1922)	$\frac{S}{\ln A}$
balance	Berger-Parker (Berger and Parker, 1970)	$\frac{N_{max}}{N}$
	Shannon Evenness (Pielou, 1969)	$\frac{-\sum p_i \ln p_i}{\ln S}$
	McIntosh Evenness (Pielou, 1969)	$\frac{N - \sqrt{\sum n_i^2}}{N - N\sqrt{S}}$
'dual concept'	Brillouin (Pielou, 1969)	$\frac{\ln N! - \sum \ln n_i!}{N}$
	Hurlbert 'rarefaction' (1971)	$\sum_i \left[1 - \frac{\binom{N-N_i}{n}}{\binom{N}{n}} \right]$
	McIntosh (1967)	$\sqrt{\sum_i n_i^2}$
	McIntosh Diversity (Pielou, 1969)	$\frac{N - \sqrt{\sum_i n_i^2}}{N - \sqrt{N}}$
	Shannon-Wiener (Shannon and Weaver, 1962)	$-\sum_i p_i \ln p_i$
	Simpson (1949)	$\sum_i p_i^2$

Figure 3. Non-parametric measures of ecological diversity. Figure extracted from [Stir98].

2.1.3.3 Diversity heuristic in Stirling's general framework for analysing diversity [Stir07]

After establishing that the three main properties for diversity are variety, balance and disparity (view section 2.1.2.3), Stirling identifies the difficulties that have to be faced when designing a

single general index/measurement of diversity. In essence, there are characteristics of diversity that are field-dependent and frustrate the possibility of creating a single general or cross-field index of diversity. Specifically, Stirling identifies the following three obstacles:

- (a) *Aggregation*: refers to the complexity of creating a single general index of diversity that can aggregate (*e.g.*, assign weights) the properties of diversity (*i.e.*, variety, balance and disparity) in a manner that is accurate across all fields.
- (b) *Accommodation*: refers to the complexity of accommodating the existing variety of perspectives on disparity in a single general index of diversity. In different fields, there are several understandings/measures of disparity that depend on the specialized criteria of the field and application context.
- (c) *Articulation*: refers to the complexity of articulating the single general index of diversity with other wider aspects of interest used in analyses or evaluation of a system. For example, “*Alongside diversity, for instance, the different species or habitats constituting ecosystems may also be assessed in terms of their conservation, agronomic, socio-cultural or aesthetic landscape qualities and values*” [Stir07]. These are important system-level properties that are clearly field-dependent. Even though these aspects are independent from diversity they might have positive or negative impacts over diversification and vice versa.

Despite the fact a single definitive general index of diversity seems to be unviable, Stirling proposes a flexible general diversity heuristic. Different “*instantiations*” of the heuristic would define specific values for its parameters (*e.g.*, weights for variety and balance) and thus behave as an index. Stirling defines 10 quality criteria or desirable features that a heuristic for diversity should have: (i) scaling of variety, (ii) monotonicity of variety, (iii) monotonicity of balance, (iv) monotonicity of disparity, (v) scaling of disparity, (vi) open accommodation, (vii) insensitivity to partitioning, (viii) parsimony of form, (ix) explicit aggregation, and (x) ready articulation. These criteria ensure that the heuristic is complete, parsimonious/simple and the principal features of diversity are explicitly considered (*i.e.*, properties —variety, balance and disparity— and obstacles —aggregation, accommodation and articulation—). Stirling specifies that *no established diversity index satisfies all these criteria*.

Stirling defines *Equation 4* as his diversity heuristic.

$$\Delta_{diversity}(\mathbb{S}) = \sum_{\substack{i=1, \\ s_i \in \mathbb{S}}}^{N-1} \sum_{\substack{j=i+1, j \neq i \\ s_j \in \mathbb{S}}}^N \left(d(s_i, s_j) \right)^\alpha \left(p(s_i) \cdot p(s_j) \right)^\beta \quad (4)$$

Equation 4. Stirling’s diversity heuristic [Stir07]

Where:

- $d(s_i, s_j)$ is the measure of disparity/difference between elements s_i and s_j (disparity).
- $p(s_x)$ is the measure of the proportional representation of elements s_x in the system (balance).
- α and β are control parameters that can take all possible permutations of values of 0 and 1. The different combinations capture various characteristics of the four properties of interest: variety, balance, disparity and diversity (view *Table 1*).

Property	α	β	$\Delta_{diversity}(\mathbb{S})$	Equivalents	Interpretation
Variety	0	0	$\sum_{\substack{i=1, \\ s_i \in \mathbb{S}}}^{N-1} \sum_{\substack{j=i+1, j \neq i, \\ s_j \in \mathbb{S}}}^N 1$	$\frac{N(N-1)}{2}$	Scaled variety
Balance	0	1	$\sum_{\substack{i=1, \\ s_i \in \mathbb{S}}}^{N-1} \sum_{\substack{j=i+1, j \neq i, \\ s_j \in \mathbb{S}}}^N (p(s_i) \cdot p(s_j))$	$\frac{\left(1 - \sum_{s_i \in \mathbb{S}}^N p(s_i)^2\right)}{2}$ Equivalent to Gini/2 or (1-Simpson)/2 (view Figure 3)	Balance-weighted variety
Disparity	1	0	$\sum_{\substack{i=1, \\ s_i \in \mathbb{S}}}^{N-1} \sum_{\substack{j=i+1, j \neq i, \\ s_j \in \mathbb{S}}}^N d(s_i, s_j)$	Un-normalized Pairwise Dissimilarity (view Equation 2)	Disparity-weighted variety
Diversity	1	1	$\sum_{\substack{i=1, \\ s_i \in \mathbb{S}}}^{N-1} \sum_{\substack{j=i+1, j \neq i, \\ s_j \in \mathbb{S}}}^N (d(s_i, s_j)) (p(s_i) \cdot p(s_j))$	$\Delta_{diversity}(\mathbb{S})$	Balance/disparity-weighted variety

Table 1. Four variants of Stirling's diversity heuristic [Stir07]

The formula is founded on the idea that diversity is “the sum of pairwise disparities, weighted in proportion to contributions of individual system elements” [Stir07]. Also, the heuristic complies with all quality criteria except for (x), which refers to the articulation of diversity with other system-level aspects of interest. To consider this criterion the author incorporates the proposed diversity heuristic in an augmented function that also takes into account other system-level properties. Further studies on this function would allow to explore the trade-offs between diversity and other system properties. For more information refer to [Stir07].

2.1.4 Dissimilarity metric

A *dissimilarity metric* measures how different or ‘distant’ two objects are from each other. The measurement of dissimilarity is fundamental to measure diversity: “if dissimilarity cannot be defined for a pair of objects, then it is difficult to imagine how collective dissimilarity can be defined for a collection of more than two objects” [Weit92]. In consequence, this work assumes a dissimilarity value can be obtained for any pair of items in order to get a diversity measurement. This value would be obtained from a dissimilarity metric that follows the properties that will be defined in this section.

The dissimilarity metric is application-dependent and is strongly related to the nature of the diversification problem [Weit92]. Weitzman indicates: “for the same collection of objects, it might be appropriate to use different distance measures in different contexts depending on the purpose” [Weit92].

The first intuition is to view dissimilarity as a distance function. A distance metric $distance(i, j)$ that measures the distance between objects i and j must meet all the following principles [Chen09]:

- (i) Non-negativity $distance(i, j) \geq 0$

- (ii) Identity of indiscernibles $distance(i, j) = 0$ if and only if $i = j$
- (iii) Symmetry $distance(i, j) = distance(j, i)$
- (iv) Triangle inequality $distance(i, j) \leq distance(i, k) + distance(k, j)$

According to Weitzman [Weit92], dissimilarity metrics meet criteria (1)-(3). He explains he does not assume the triangle inequality holds for dissimilarity metrics in order to cover nonmetric distances. In essence, what is important is that for every pair of elements there is a nonnegative symmetric measure that expresses the dissimilarity between the objects.

Another way to view dissimilarity or a distance function is as the contrary of similarity. Similarity helps to quantify the mutual information shared between two objects [Chen09]. Chen *et al.* [Chen09] define for a similarity metric $sim(x, y)$, which measures the similarity between objects x and y , the following rules:

- (i) $sim(x, y) = sim(y, x)$
- (ii) $sim(x, x) \geq 0$
- (iii) $sim(x, x) \geq sim(x, y)$
- (iv) $sim(x, y) + sim(y, z) \leq sim(x, z) + sim(y, y)$
- (v) $sim(x, x) = s(y, y) = s(x, y)$ if and only if $x = y$

Chen *et al.* [Chen09] also propose two functions used to transform a similarity metric $sim(x, y)$ into a distance metric. These transformations can be viewed in *Equation 5*.

$$F_p(sim(x, y)) = \frac{sim(x, x) + s(y, y)}{2} - s(x, y) \quad (5)$$

$$F_m(sim(x, y)) = \max\{sim(x, x), sim(y, y)\} - sim(x, y)$$

Equation 5. Transformations of similarity metric into distance metric [Chen09]

For example, if we define the maximum similarity as 1 and thus $sim(x, x) = 1$ and $sim(y, y) = 1$, it can be seen that for both transformations *Equation 6* emerges.

$$distance(i, j) = 1 - sim(i, j) \quad (6)$$

Equation 6. Distance as the inverse of similarity

Where $distance(i, j)$ is the distance between objects i and j , and $sim(i, j)$ the similarity between the same objects. This last definition is very intuitive and used in several applications such as [McSh02].

2.1.5 Diversification Problem

We refer to the process of creating a diverse set of elements as *diversification*.

The diversification problem can be expressed as: The task of selecting a subset of k elements from a broader set \mathbb{S} in order to maximize an objective function that considers both the *quality* and *diversity* of the k elements.

The criteria used to define the quality of each item are application dependent. The challenge is that there usually is a trade-off between diversity and quality if the broader set \mathbb{S} contains elements that are similar from each other, in other words, if there is redundancy within the elements in the set. This means if one element is of high quality then all similar elements have high quality as well. However, all similar elements cannot be added because this would reduce

diversity. The challenge is then how to choose k elements in order to both maximize quality and diversity at the same time.

The following sections will discuss the diversification problem for both the areas of Information Retrieval and Recommender Systems. Given that the diversification problem in the field of Information Retrieval serves as foundation for research on diversification in Recommender Systems, it is important to highlight the associations and differences between the characterizations of diversity from both areas, showing that the problem definition is related but different.

2.2 Diversity in Information Retrieval

A simplified definition of the *Information Retrieval (IR)* task is to identify an ordered set of relevant information resources (or documents) that respond to an explicit or implicit user information need. On the one hand, an explicit user need is directly expressed by the user, usually in the form of a query. On the other hand, an implicit user need can be inferred from the User Profile, which contains information about past interactions of the user with the system. *IR* systems help decrease user Information Overload by reducing the document search space users need to explore to meet their information need. A type of *IR* system are *Recommender Systems (RecSys)*, which respond to implicit user needs in order to promote product discovery. *RecSys* will be discussed in *section 2.3*. In this chapter, *IR* systems that receive an explicit query input will be analyzed.

Extensive work on the field of *IR* has been carried out to solidify concepts related to the diversification problem [Varg12]. In this manner, the study of diversity as it has been applied in *IR* serves as a strong foundation for work on diversity in *RecSys*. The aim of this section is to introduce the diversification problem in *IR*. In first place, *section 2.2.1* defines *IR* within the context of the present research. In *section 2.2.2* several motivations are presented to show that diversity is a desirable feature of *IR* systems. Lastly, *section 2.2.3* formalizes the Diversification Problem in *IR*.

2.2.1 Defining Information Retrieval

As has been mentioned, the goal of an *IR* system is to provide a user an ordered set of relevant documents that satisfies the user's information need. To accomplish this, an *IR* system needs to predict/estimate which documents a user would find relevant/useful [Gord92]. A traditional approach is to measure a document's quality or relevance by considering the similarity of the document to the user query. Retrieved documents would be those with the highest relevance scores, and are ordered according to the *Probability Ranking Principle*. This principle indicates that the optimal rank is obtained by retrieving documents in decreasing order of their probability of relevance (*i.e.*, relevance score) [Maro60][Robe77][Carb98][Gord92].

Even though new and different approaches for *IR* have been proposed, this strategy is the closest to *RecSys* techniques and therefore the most significant to our research. Correspondingly, in traditional *RecSys* the utility of an item is estimated as a rating (analogous to the relevance score) and recommendations are offered to users in decreasing order of predicted rating (analogous to the probability ranking principle). More on *RecSys* can be viewed in *section 2.3*.

The following section will analyze why considering diversity explicitly for *IR* is a desirable feature.

2.2.2 *Motivation: Why consider diversity for Information Retrieval?*

Recent research has emphasized the need for *IR* systems to provide a diversified set of results [Ceri13]. However, *IR* systems do not offer diversified results naturally, especially when the document space contains highly redundant/duplicative information [Carb98][Ceri13]. As this is mostly the case for *IR* systems, documents that are similar to the target query also tend to be similar to each other [McSh02]. As a consequence, *IR* systems usually provide users an over-specialized homogenous set of results. This is not desirable, it would mean that if one document is not relevant then all similar results are not relevant, and in consequence, there is a high risk of not satisfying the user [Ceri13].

Another reason why *IR* systems do not offer diverse results naturally is because the relevance of a document is calculated independently of other documents with which it will co-exist in the result list [Robe77]. Independent assessment of relevance is a key feature for the *Probability Ranking Principle* to hold true [Gord92]. However, considering the interdependent nature of the value of documents could have a positive impact on the perceived quality of results [Radl09][Agra09][Sant10]. In utility theory, the *law of diminishing returns/marginal utility* explains that the perceived utility/usefulness/satisfaction of a product to be consumed depends on the products that have been consumed in the past [Mank04]. This law represents the notion that there is decrease in value of a product when its consumption increases (*i.e.*, more is better... up to a point) [Mank04]. As a result, a rational consumer would seek to maximize his/her total utility by “*purchasing a combination of different products rather than more of one particular product*” [Mank04]. In the context of *IR*, the *law of diminishing returns* indicates that higher diversity within the result set would lead to an increase of the overall usefulness of results.

Diversity in *IR* should also be considered as a response to the ambiguity of user queries and intents [Ceri13][Agra09][Sant10]. Because the same search engine must respond to several users at a time, it has to satisfy users that have varied needs, intentions and that place different meanings to the same query [Akin12]. Also, it has been found that a great fraction of queries are short thus tend to be under-specified [Akin12][Zhen12][Goll09]. In these circumstances, using the *Probability Ranking Principle* and ordering results according to relevance could produce sub-optimal rankings [Goll09]. Without further information to help determine the precise user intent at a specific context, an *IR* system should collect results that respond to the varied interpretations of the query within a top-*k* list and hope that as many users as possible find at least one document that responds to their information need [Agra09][Akin12][Goll09]. Even though results could be focused only on the “*best*” interpretation of the query, a diversified list of results with high coverage of the different interpretations or sub-topics of the query is less risky, and could increase the chance of satisfying a user with a random intent [Ceri13][GilC13][Goll09][Zhen12]. This is a similar effect as in *Modern Portfolio Theory* (view *section 2.1.1*), a restricted set of choices in the result set infers a higher risk of not satisfy the user given the uncertainty surrounding the user need, analogous to the uncertainty of stock returns.

From a related perspective, providing diversity could increase the amount of information transmitted by the result set and in turn increase the chance of responding the user need.

Similarity can be seen as the measure of the common information/features shared by two objects. From an information theory perspective, if each object in the set is considered as an event, then a more uncertain/unique event will offer more information. In this manner, to offer users the most information within the result set, diversity should be added to promote both uniqueness and usefulness of the information provided: the value of an object does not only depend on its uniqueness but if it represents higher-valued features [Nehr02].

In general, the main reason to consider diversity for *IR* is to augment the probability the user will find a relevant document within the retrieved set of results, especially in face of high uncertainty associated to the user query. The purpose of diversity is to remove redundancy from the set and offer more options and a broader array of choices. The goal of diversification would be to select documents that are not only relevant to the query but that also cover as many query interpretations as possible considering the interdependence between the documents in the result list.

2.2.3 Problem Definition: Diversification Problem in Information Retrieval

In this section, the *IR* task is linked to the Diversification Problem discussed in *section 2.1.5*. Correspondingly, *IR* systems aim to maximize an objective function that considers both the quality (*i.e.*, relevance) of the information items retrieved and the diversity of the items within the result set. However, there is a clear trade-off between selecting items that are of higher relevance (which as explained in *section 2.2.2* tend to be similar to each other) and obtaining diverse results [Goll09]. Therefore, the diversification problem in *IR* is usually modelled as a bi-criteria optimization problem that aims to find the appropriate trade-off balance between two competing objectives: maximizing diversity and maximizing relevance [Goll09].

One way to specify the diversification problem in *IR* is as follows: Let \mathbb{Q} be an ambiguous query associated to diverse query intents or sub-topics s (*i.e.*, $\mathbb{Q} = \{s_1, s_2 \dots, s_n\}$). Let \mathbb{D} be a corpus of documents such that every document $\mathbb{d} \in \mathbb{D}$ is also associated to various sub-topics (*i.e.*, $\mathbb{d} = \{s_1, s_2 \dots, s_m\}$). A document is relevant to a query if the function $rel(\mathbb{d}, \mathbb{Q})$ is high. The diversification problem aims to find a subset of documents $\mathbb{R} \subseteq \mathbb{D}$ of size k (*i.e.*, $|\mathbb{R}| = k$) which together cover as many sub-topics that can be found in \mathbb{Q} and that are also maximally relevant to \mathbb{Q} . In [Agra09][Cart11][Goll09][Sant10], it is demonstrated that this problem can be reduced to the *Maximum Coverage Problem* which is an *NP*-hard problem related to the *Set Cover* problem. In the *Maximum Coverage Problem*: “one is given a universe of elements \mathcal{U} , a collection \mathcal{C} of subsets of \mathcal{U} , and an integer k . The objective is to find a set of subsets $\mathcal{S} \subseteq \mathcal{C}$, $|\mathcal{S}| \leq k$, to maximize the number of covered elements” [Agra09].

Though the problem is *NP*-hard it can be solved with a greedy heuristic. Agrawal *et al.* [Agra09] point out the work of Nemhauser *et al.* [Nemh78], which demonstrate that a greedy strategy aiming to maximize a monotonic (*i.e.*, non-decreasing) submodular function² can achieve an approximate solution with a bounded error. This bounded error is defined in *Equation 7*, where: $z(S)$ is real-valued submodular function, S is a subset of the larger set N , K is the maximum cardinality that set S can have and e is the base of the natural logarithm. In order to take advantage of this bound, Agrawal *et al.* [Agra09] demonstrate their optimization function (view *section 3.1.1.2*) is submodular. Furthermore, Agrawal *et al.* [Agra09] highlight

² Refer to [Nemh78] for a definition of submodular function.

that a submodular function naturally accounts for the *law of diminishing returns*. This is due to the fact that in a submodular function as the set increases it will be harder for an object to add to the marginal benefit of the set.

$$\max_{S \subseteq N} \{z(S) : |S| \leq k, z(S) \text{ submodular}\} \quad (7a)$$

$$\frac{\text{value of greedy approximation}}{\text{value of optimal solution}} \geq 1 - \left(\frac{K-1}{K}\right)^K \geq \frac{e-1}{e} \quad (7b)$$

Equation 7. Nemhauser bound for greedy optimization of a submodular function [Nemh78]

2.3 Diversity in Recommendation Systems

This section introduces the main field of this research which is *Recommendation Systems (RecSys)*. In first place, *section 2.3.1* defines what a *RecSys* is and highlights the difference between *RecSys* and *IR*. Next, in *section 2.3.2*, it is explained that diversity in *RecSys* is a desired feature. However, in *section 2.3.3*, it is argued that *RecSys* do not offer diversified results naturally and the diversification problem in *RecSys* is formalized.

2.3.1 Defining Recommendation Systems

Recommender Systems (*RecSys*) have emerged as tools that help users easily identify worthwhile and interesting products by means of proactive personalized suggestions. These suggestions guide users in exploring a large product space by pointing out the path to potentially useful products they might not have been intentionally looking for or might not have even known existed.

Specifically, a recommendation is a set of N items ordered to maximize the utility or value of items that are at the top of the list [Adom05]. In the most common setting, utility is determined by a *rating* or score of numeric value that represents how much a particular user likes/dislikes a determined item. User rating information is captured in the User-Item Matrix (view *Figure 4*).

		Items					
		1	2	...	i	...	m
Users	1	5	3		1	2	
	2		2				4
	:			5			
	u	3	4		2	1	
	:					4	
	n			3	?		2

Figure 4. User-Item Matrix in Recommendation Systems

Ideally, ratings would be assigned explicitly by users. However, given the large size of most item spaces, it is improbable that users can or are willing to assess all possible items. In consequence, the rating score for most items is generally missing, making the User-Item Matrix very sparse. Hence, the recommendation problem is centered on the prediction of the score/utility that the user would assign to an unrated product.

Formally defined, the Recommendation System's main challenge is: "Given a set of users (or customers) C and a set of items (or services) S with or without associated user/item features and a set of interactions between users and items (users' ratings on items or transactions of

user-item pairs), *predict the exact or relative utility of individual items for individual users*” [Adom08].

After extrapolating utility values to the whole *CXS* space, *i.e.*, once an estimated value is computed for each one of the nonrated items, a Recommendation System can then determine the ordered list of products that maximizes a user’s overall utility. Analogous to *IR* and founded on the *Probability Ranking Principle*, the optimal rank in *RecSys* is obtained by recommending items in decreasing order of their predicted utility value. There are several types of *RecSys*, generally classified keeping in mind their estimation techniques and the information taken into account when manufacturing a recommendation. These will be explored in *section 2.3.1.2*. The following section clarifies the difference between *IR* and *RecSys*.

2.3.1.1 Recommendation Systems vs. Information Retrieval

The main difference between Recommendation Systems and Information Retrieval Systems lies on the difference between searching and browsing. On the one hand, *IR* systems are designed to aid the task of searching by responding to an explicit user information need. On the other hand, *RecSys* are tools that guide users towards interesting products without the need of a specific information need, and therefore aid the task of browsing.

Searching and browsing are two related but different concepts. When searching a user has a set of well-defined objectives which are clearly identified before initiating the search process [Morv98]. Alternatively, when browsing, a user has unclear goals or vaguely defined information needs, and navigates through results based on her/his interests, which may change during the interaction with the system [Morv98]. It is important to highlight that browsing encourages discovery, allowing users to stumble on something new which was completely unknown to them. Browsing and searching are not mutually exclusive; sometimes users choose to switch between searching for something specific to casual browsing when exploring results [Morv98]. In the book “*Information Architecture for the World Wide Web*” [Morv98], while explaining browsing and searching, the author states: “*If you care about the consumer, make sure your architecture supports both modes.*”

IR systems are established under the assumption that a user is able to accurately portray her/his information needs in the form of a search query. Nevertheless, users generally have an unclear idea of what they want or are unable to explain exactly what they are looking for. Furthermore, occasionally users do not even know what they want at all; they just want to receive interesting information within their context of use. In these use cases, a better and more suitable solution is offered by *Recommendation Systems*.

Recommendation Systems have emerged to guide users in the task of efficiently browsing/navigating a large product space of alternatives even when the user has unclear information needs (*e.g.*, I want a recipe that has chicken, what do you recommend?) or none at all. *RecSys* can respond proactively to implicit user needs, especially those the user is not aware of, by suggesting novel products/services to a user that she/he didn’t even know existed, much less wanted, helping the user to discover in less time something completely new he/she might not have found on their own. The heart and value of *RecSys* lies on notions such as “surprisal”, “unexpectedness”, novelty, discovery and serendipity. Jeffrey O’Brien from Fortune magazine explains discovery in the following phrase: “*The web, they say, is leaving the era of search and entering one of discovery. What’s the difference? Search is what you do when*

you're looking for something. Discovery is when something wonderful that you didn't know existed, or didn't know how to ask for, finds you”.

Essentially, the most important difference between Recommendation Systems and traditional *IR* systems, is that the second allows users just to search, while Recommender Systems enrich searching and allow users to browse and discover.

2.3.1.2 Types of Recommender Systems

As was mentioned in *section 2.3.1*, the *RecSys* problem can be reduced to an estimation strategy that can predict the utility score for unknown ratings. As a result, *Recommender Systems* are usually classified according to the distinct strategies used to achieve this task.

According to Adomavicius *et al.* [Adom08], *Traditional Recommender Systems* consider only information about users and items to estimate a single value rating that a user would give to an item. Alternatively, non-traditional *RecSys* would alter some of the views of traditional approaches, for example, consider contextual features or use multi-criteria ratings. Within this research, we will focus on traditional *RecSys* only, as non-traditional *RecSys* are out of the scope of the project.

Traditional Recommendation Systems are based on three techniques [Adom08]: (a) *Content-based*: the system recommends items similar in content to those items the user has liked in the past, based on product descriptions and the known ratings of a particular user [Adom05], (b) *Collaborative Filtering*: a user is recommended items based on known ratings of other users [Adom05] and, (c) *Hybrid approaches*: these methods combine collaborative and content-based methods.

On the one hand, Content-based approaches are based on the idea that users will like items similar in content to those he has liked in the past. These approaches use the information within *item profiles* and *user profiles* in order to estimate ratings. An *item profile* describes an item in terms of features/attributes, and how much each feature is representative of an item's content. A *user profile* describes the user's preferences, and how important each preference is for the user. Each time a user rates an item, the user preferences are updated considering the new rating and the item's profile. In order to estimate the rating for an unrated item, a matching function compares the item's profile and the user's profile to generate a score for the item. The easiest approach to achieve this is to draw a one-to-one relationship between item features and user preferences (*e.g.*, in a news story recommender both item preferences and features would be keywords from the stories). In this case, both item and user profiles can be represented by weighted feature vectors, and the relationship between both can be calculated using a similarity metric such as cosine similarity.

On the other hand, Collaborative Filtering (*CF*) only uses the information in the User-Item Matrix (view *Figure 4*). These approaches are divided in two strategies: user-based and item-based. In both the underlying heuristic is that similar users tend to rate items in a similar fashion and will continue to do so in the future.

In user-based *CF*, item recommendations for a target user are generated based on the information of similar users. In this case, a user profile contains the ratings that a user has explicitly given to items. Two users are similar if their profiles are similar, *i.e.*, the users have rated items in a similar fashion. To retrieve item recommendations, the *k* most similar users to the target user are identified and item ratings are predicted for candidate items, which are items the neighborhood has rated but the user has not. The estimated rating for a candidate item is

obtained from an aggregate of the ratings from the users in the neighborhood who have rated the candidate item. Candidate items with the highest predicted rating are recommended to the user.

Another perspective is item-based *CF*, where the user's predicted rating for an item is calculated according to similar items. In this case, an item profile contains the ratings that users have given to the particular item. Two items are similar if their item profiles are similar, *i.e.*, users tend to rate both items in a similar fashion. To predict the rating a user would give to a target item, the k most similar neighbors to the target item are found. Next, an aggregate of the ratings that the user has given to items in the neighborhood is calculated to estimate the rating for the target item.

Recommendations generated with traditional techniques usually don't produce diverse results. However, diversity is a desirable feature for a Recommender System which ideally should suggest to users a list of varied products that at the same time maximizes the overall utility perceived by the user [Mcne06][Adom05]. An analysis of why diversity is a desirable feature for *RecSys* will be provided in the following section.

2.3.2 Motivation: Why consider diversity for Recommender Systems?

In this section two perspectives are presented that explain why diversity is important for *RecSys*: *section 2.3.2.1* argues diversity is important to deal with the uncertainty surrounding the user profile, and *section 2.3.2.2* discusses diversity is essential to the concept of novelty which in turn is fundamental to *RecSys*.

2.3.2.1 On the uncertainty of the User Profile

The only evidence of user tastes a Recommender System has is encapsulated within the user profile. However, much like a user query in *IR*, the user profile is incomplete and ambiguous. This can be explained by the large size of item spaces and the unfeasibility of obtaining explicit rating information on all products from users, the unreliability of interpreting implicit information to understand user tastes and the dynamic nature of user preferences. As has been explained, diversity is a good solution in face of uncertainty. A *RecSys* should offer users a diverse set of suggestions representative of the variety of the user's tastes in order to increase the chances the user finds useful items within the recommendations [Varg12][Zhen12].

As in *IR*, where diversity aims to cover as many of the sub-topics that are associated to the query as possible, *RecSys* should aim to cover as many of the user preferences as possible. Ziegler *et al.* [Zieg05] present the following example: in a book recommender system a client that has purchased various books of a particular author may obtain a recommendation list of other books of the same author (homogeneous recommendation list). If we consider that the user has other interests, for example, history novels or cuisine books, the recommendation list seems poor and incomplete. Surely the user would be able to discover other books of an author he/she has been previously interested in on his/her own. If this is so, the recommendation would have no effect on the user's behavior or represent major utility for the user. In addition, because the items in the recommendation list are similar to each other, if one is not useful then all of them are not useful. Thus, a logical assumption is that users perceive higher quality/utility from a recommendation list when they are presented a wide range of options that cover most of their overall preferences, and not a homogeneous/monotonous set of alternatives associated to few or one preference.

2.3.2.2 *Serendipity, Novelty and Diversity*

In first place, let's define the concepts of serendipity and novelty and their relation to diversity. In their seminal paper on evaluation of *RecSys*, Herlocker *et al.* [Herl04] clarify the difference between *novelty* and *serendipity*. Novelty is referred to items that the user is not aware of or hasn't seen before. Serendipity requires a stronger notion of novelty, and is defined as an item both novel and unexpected or non-obvious. An obvious recommendation refers to a product the user could have found on their own, without the aid of the recommender; for example a highly popular product or an item too similar to what the user has already seen (*e.g.*, a book from an author the user already knows and has read many books on). In a nutshell, an item can be novel but not serendipitous if it is unknown to the user but he/she could have eventually discovered it. However, in essence, *serendipity* is just a stronger view of novelty. This leads to a view of novelty as a gradual measure, where different levels of *novelty* can be achieved depending on how far or *diverse* an item is from the user's past experience. The only direct evidence of past experience is given by the user profile. In this manner, a novel but obvious item could be viewed as "different" from the user profile but not as "different" as a serendipitous item is. Nonetheless, considering only the user profile to define novelty is not enough as the user's past experience could be defined by other factors, such as product popularity. In order to encourage discovery of new products it is generally desired to suggest items that belong to niches (less popular), given that it is more likely that the popular products are already known to the user or the user can discover the product on their own. Vargas [Varg11] define novelty as "*the difference between an item and "what has been observed" in some context*". They offer various examples of what context could be: a group of users, past or alternative recommendations, advertisements, among others.

To summarize, diversity under the notion of novelty is a highly desirable feature: it's directly related to the idea of discovery and fundamental to the purpose of *RecSys*. As an additional advantage, diverse/novel recommendations help increase the information flow between the user and the system. It is to be expected that discovering new products would lead to an information gain for the user but this is also true for the *RecSys* itself. Discovery of new items leads to user feedback on diverse/novel items. This feedback generates larger information gain for the user profile than feedback of non-novel items, broadening the knowledge over the user preferences [Lemi08].

2.3.3 *Definition of research problem: Diversification Problem in Recommendation Systems*

Traditional Recommendation Systems do not offer diverse recommendations naturally. The reasons are related to the heuristics of traditional recommendation techniques and current evaluation metrics, which ignore diversity and evaluate a recommendation list as an aggregate of the individual scores of items and not as a complete entity. In detail, the reasons are as follows:

- (a) *The heuristics that lay foundation to recommendation techniques are based on similarity measures that limit the diversity of items considered for recommendation.*

It is improbable for a *RecSys* to offer diverse product suggestions when using traditional recommendation techniques that are centered on similarity-based heuristics. This problem can lead to the perception that a user only sees items which are similar to what the user has seen

before, which is commonly referred to as the “*filter-bubble problem*”. In a nutshell, this is due to (i) overspecialization, (ii) bias towards popular items, and (iii) bias towards items which are similar to highly-rated items from the user profile.

In first place, content-based recommendation techniques suffer the problem of overspecialization. By definition these techniques only consider products that are similar to those that the user has evaluated in the past [Adom05]. This indicates that the *RecSys* is not in the capacity of suggesting products that are too different from products previously preferred by the user. As a consequence, content-based recommendations tend to offer a set of suggestions that lack both diversity and novelty.

In second place, collaborative-filtering techniques also produce non-diverse recommendations, even though they are less susceptible to this problem than content-based methods. First off, collaborative-filtering techniques overall tend to be biased towards the most popular products (those that the majority of users have liked the most). Nevertheless, a suggestion for a less popular but interesting product can be more valuable for the user [Ande06b]. Particularly, McNee *et al.* [Mcne06] indicate that Item-Item collaborative filtering algorithms can trap users in a ‘*similarity hole*’ offering suggestions only of products that are very similar to each other. They state that “*this problem is more noticeable when there is less data on which to base recommendations, such as for new users to a system*” [Mcne06]. In the same way, in User-User collaborative filtering, if user profiles are too similar to each other they could have ratings for the same narrow group of items, which limits the diversity of *RecSys* results [Smyt01][Said12][Yang13][Zhan09].

Lastly, hybrid techniques combine content-based and collaborative-filtering techniques in order to mitigate the problems of one with the advantages of the other. With hybrid techniques more diversity on results can be obtained, however it is likely that recommendation results could also be limited to a certain space in the product catalog: popular products and products which are similar to those that the user has rated positively in the past.

(b) *Evaluation metrics that assess the individual quality of items in Recommendation Systems penalize diversity and novelty.*

Currently, most publications on the topic of *Recommender System Evaluation* focus on accuracy metrics. In general, these metrics measure the ability of the system to predict the utility/rating that a user would assign to a product that has not been rated yet [Herl04]. Nevertheless, additional metrics should be considered to measure true user satisfaction given that high accuracy alone does not always indicate users will obtain interesting product suggestions that are useful for practical purposes [Chen13][Herl04][Mcne06][Adom08][Zieg05]. For example, in a supermarket application the system may suggest products that the user will most probably buy regardless of the recommendation. To be exact, obvious and easy to predict recommendations, such as milk and bread [Herl04][Adom08]. These suggestions are not useful to the user, as they would not change the user’s behavior (*i.e.*, the user had previously planned to purchase the items independently of the suggestion). However, these suggestions will achieve a high level of accuracy given that users usually like and buy these products regularly. Furthermore, because a supermarket system would allow repetitive consumption (and therefore recommendations) of products; the next time the user receives a recommendation they could receive the same exact suggestion of products. It would be much more valuable to suggest to the user a new product, for instance in the area of frozen meals, that the user does not know about and that the user could potentially love [Herl04]. The problem is that with accuracy

metrics a novel product (such as the frozen meal) would receive a lower predicted rating compared to products similar to those that the user always consumes (e.g., bread and milk), precisely because it is different. As can be seen, there is a tradeoff between accuracy and diversity of product recommendations. This shows, that in addition to accuracy, other metrics such as *novelty* and *diversity* should also be considered to obtain better insight on real user satisfaction [Varg11][Herl04].

In general, it is important to take into account that Recommendation Systems should be evaluated with respect to actual user satisfaction and not just pure accuracy [Zieg05]. It is logical to think that the utility perceived by the user for a recommendation is higher for less obvious items the user could be interested in, than a recommendation of products very similar to or the same as products that the user has viewed/consumed in the past [Zhan08]. Rather than receiving predictable and monotonous suggestions of the same type of products, the user would receive interesting and novel product recommendations leading the user to discover something new and useful [Hurl11]. Mcnee *et al.* [Mcne06] sustain that a user measures his/her level of satisfaction with the results of a Recommender System considering the ability of the recommender to suggest products that the user would not have thought of or found on his/her own.

(c) *Recommendation list evaluation is performed as an aggregate of the individual scores of items, disregarding the real value of items in the context of the list.*

In order to evaluate the quality of a recommendation list, current practices aggregate the individual accuracy scores of the products in the list. However, the list should be assessed as a whole and the quality of each product under the context of the list [Mcne06]. For example, if the first item in the list is a Star Trek movie it is not necessary that the second and third item in the list also be of Star Trek related products (law of diminishing returns). However, if the user is a big fan of Star Trek and each product is evaluated independently, the system senses that the recommendation as a whole has a high level of accuracy (and therefore quality) if all products are Star Trek related. Adding to the list a non-Star Trek product that might have slightly lower predicted rating but add more information/value to the list is penalized by current list metrics. In short, recommendation list metrics do not evaluate each product within the context of the list and cannot determine if the list offers items that are both of high quality and sufficiently diverse to cover the spectrum of the user's interests [Mcne06].

In brief, traditional recommendation techniques and evaluation metrics are not enough to achieve diversity-aware Recommendation Systems. It can also be seen that the diversification problem in *RecSys* is similar to that in *IR*, where there is a clear trade-off between the individual accuracy of an item and the overall diversity of the recommendation list. In this manner, the diversification goal in *RecSys* is to generate a list of suggested items that maximize both the predicted rating for items and coverage over the wide spectrum of user preferences (equivalent to covering as many of the query interpretations as possible in *IR*) [Chen13]. But this is not the complete picture for diversification in *RecSys*. Recommendation results must also account for novel products, which by definition are not directly related to the user preferences, in order to encourage product discovery. This brings up an additional trade-off between how much the *RecSys* wants to *exploit* the known information about the user by covering the preferences in the user profile, and how much the *RecSys* wants to *explore* what other preferences the user could have by offering novel products.

The trade-off between exploitation and exploration could depend on factors such as the maturity of the user profile and the user's openness to experience. If the user is a new user, and there is little information on his preferences in the user profile, then it would be logical for the *RecSys* to attempt to show more novel products in order to gain information about the user's interests. However, if the user is new to the system, it has been found that less novel and more familiar recommendations help build the user's trust of the *RecSys* [Herl04]. In addition, users that are less open to new experiences would prefer to receive recommendations of products similar to those they have liked in the past. On the other hand, users that are more open to experience would prefer novel recommendations that introduce them to unexplored areas of the product catalog [Tint13]. Openness to experience could depend on the product domain, and the risk/cost associated with consuming and not liking a product. McSherry [McSh02] explains there are domains where adding diversity at the expense of showing all relevant even though similar items is not advisable, for example: "*when the recommended items (e.g. jobs, rental apartments, bargain holidays) are limited in number, available for a limited period, or sought in competition with other users.*"

In conclusion, this leads to the definition of the research problem which we will address in this research report: Design a diversification technique for Recommendation Systems that generates a diversified list of results that balances the trade-off between quality (in terms of accuracy) and diversity considering the trade-off between exploitation of the user profile and exploration of novel products.

2.4 Summary

In this chapter we have laid the conceptual foundations needed to understand the motivation and problem definition for diversity-aware Recommendation System.

In first place, we have explored the concept of diversity and found it depends on three fundamental attributes: variety, balance and disparity. In addition, we found a metric that can measure the diversity of a set in terms of the three attributes. Most importantly, the diversification problem was defined as the challenge of finding a set of elements that maximizes the trade-off between quality and diversity.

After this, the diversification problem was specified for the area of Information Retrieval. This is because Information Retrieval is a much more mature area than Recommendation Systems, and knowledge from this field can be easily transferred towards Recommendation Systems. It was found that adding diversity to a set of results in Information Retrieval involves a trade-off between the individual relevance of documents and attempting to cover as many of the query sub-topics as possible in the result list. Furthermore, the challenge of obtaining a diversified set is reduced to an *NP*-hard problem. Nevertheless, it has been proven in mathematical theory that this particular problem can be solved using a greedy approximation that under given conditions has a bounded error.

Finally, the diversification problem is defined for the area of Recommendation Systems. In first place, it was highlighted that while Information Retrieval systems allow users to search a space of documents in response to an information need, Recommendation Systems are tools created to guide users in browsing tasks oriented towards the discovery of interesting products without the need of an explicit user requirement. This clarification is important as it marks the difference between diversification approaches in the field of Information Retrieval and in the field of Recommendation Systems. Next, it is explained that Recommendation Systems do not

offer diverse results naturally because: (a) the heuristics that lay foundation to recommendation techniques are based on similarity measures that limit the diversity of items considered for recommendation, (b) evaluation metrics that assess the individual quality of items in Recommendation Systems penalize diversity and novelty, and (c) recommendation list evaluation is performed as an aggregate of the individual scores of items, disregarding the real value of items in the context of the list. Finally, the diversification problem in Recommendation Systems and main research problem of this research report is defined as not only balancing the trade-off between diversity and the individual quality of products, but doing so within the context of the trade-off between exploitation of the user profile and exploration of novel products.

In the following section, a literature review of related works on diversification techniques for both Information Retrieval and Recommendation Systems is presented. In addition, given that it was found that current evaluation metrics for Recommendation Systems are not adequate to assess diversity; current works that evaluate diversity in the field of Recommendation Systems are also discussed.

Chapter III

LITERATURE REVIEW

This chapter aims to provide a review of prior works that have addressed the diversification problem both for Information Retrieval and Recommendation Systems. In first place, diversification techniques proposed for both Information Retrieval and Recommendation Systems are presented in *section 3.1*. A comparative analysis of the works is discussed in *section 3.2*. Given that traditional evaluation metrics for Recommendation Systems do not explicitly consider diversity, an analysis of diversity-aware evaluation metrics proposed by related works is offered in *section 3.3*. Finally, a summary of findings is presented in *section 3.5*.

3.1 Diversification Techniques

This section first reviews diversification techniques for Information Retrieval in *section 3.1.1*. Next, diversification techniques for the area of Recommendation Systems are revised in *section 3.1.2*. A clear connection between diversification techniques for *IR* and for *RecSys* can be viewed, as many of the works in *IR* have served as inspiration for the proposed diversification techniques for *RecSys*. Finally, a comparative analysis of the reviewed works is shown in *section 3.2*.

3.1.1 Diversification Techniques for Information Retrieval

This section reviews diversification techniques that have been proposed for the area of Information Retrieval. All the reviewed techniques in this section are based on a greedy optimization approach, which is used to choose from a broader set of documents a subset that maximizes an objective function. In [Zhen12] the greedy optimization algorithm is defined as follows: the algorithm starts with an empty set $\mathbb{S} = \emptyset$. Then, it iteratively selects a local optimal document to add to \mathbb{S} from a broader set \mathbb{D} that satisfies *Equation 8*. This process is repeated until $|\mathbb{S}| = k$.

$$d^* = \arg \max_{d \in \mathbb{D} \setminus \mathbb{S}} (G(\mathbb{S} \cup \{d\}) - G(\mathbb{S})) \quad (8)$$

Equation 8. Greedy Algorithm Optimization Function [Zhen12]

The analyzed works in essence differ on the definition of the objective function, which is why in the reviewed works special focus will be given on the different objective functions defined.

3.1.1.1 Maximal Marginal Relevance (MMR) [Carb98]

Maximal Marginal Relevance (MMR) is a seminal method proposed by Carbonell and Goldstein in [Carb98], focused on maximizing both *query-relevance* and *information-novelty* for automatic document summarization. The objective is to re-rank a list of documents retrieved from an *IR* system in order to minimize the redundancy as well as maintain results that are relevant to the query. *MMR* is appropriate for cases where potential relevant documents are highly redundant, even containing duplicative information.

MMR aims to maximize “*relevant novelty*”, which is measured by a weighted linear combination of the independent measures of *relevance* and *novelty*: (i) a document is *relevant* if it is similar to the query and, (ii) a document is *novel* if it is dissimilar from previously selected documents. This linear combination is called “*marginal relevance*”. *MMR* takes as input a ranked list of documents \mathbb{R} retrieved from an *IR* system given the query Q . The method iteratively constructs a list \mathbb{S} that selects at each step a document from $\mathbb{R} \setminus \mathbb{S}$ with the highest marginal relevance. *MMR* is defined as in *Equation 9*.

$$MMR(Q, \mathbb{R}, \mathbb{S}) \stackrel{\text{def}}{=} \arg \max_{D_i \in \mathbb{R} \setminus \mathbb{S}} \left(\lambda \cdot \text{sim}_1(D_i, Q) - (1 - \lambda) \cdot \max_{D_j \in \mathbb{S}} (\text{sim}_2(D_i, D_j)) \right) \quad (9)$$

Equation 9. Maximal Marginal Relevance

Where the definition of each variable can be found in *Table 2*.

Variable	Definition
Q	Query or User Profile.
\mathbb{R}	Ranked list of documents retrieved by an <i>IR</i> system.
\mathbb{S}	Subset of documents in \mathbb{R} that have been selected/retrieved to be provided to the user.
$\mathbb{R} \setminus \mathbb{S}$	Subset of documents in \mathbb{R} that have not been offered to the user. Set difference between \mathbb{R} and \mathbb{S} .
$\text{sim}_1(D_i, Q)$	Function to measure how relevant/similar a document is to the query.
$\text{sim}_2(D_i, D_j)$	Function to measure similarity between two documents.
$\max_{D_j \in \mathbb{S}} (\text{sim}_2(D_i, D_j))$	The novelty an item D_i would add to the set \mathbb{S} , is given by the minimum distance (maximum similarity) between the item and the set.
λ	User tunable parameter that controls the trade-off between relevance and novelty. If $\lambda = 1$, then a standard relevance-ranked list is obtained. If $\lambda = 0$, then a maximally diverse list of documents \mathbb{S} is chosen from \mathbb{R} .

Table 2. Maximal Marginal Relevance variable definitions

Through experimentation and user studies, Carbonell and Goldstein [Carb98] found that starting with a more diverse list (*e.g.*, $\lambda = 0.3$) was an effective strategy to understand the information space. After this, if the query is reformulated given relevance feedback, raising the value of the user tunable parameter would help to focus on specific more important documents relevant to the new less ambiguous query. Carbonell and Goldstein carried out user studies with five users and found: “*The majority of people said they preferred the method which gave in their opinion the most broad and interesting topics (MMR)*” [Carb98].

3.1.1.2 IA-Select algorithm for diversifying search results [Agra09]

Agrawal *et al.* [Agra09] address the challenge of responding to ambiguous user queries in a web environment where a taxonomy of information has been defined. They propose the *IA-Select* algorithm, which diversifies search results in order to minimize the risk that the average user will be dissatisfied with search results. In addition, they adapt several classic *IR* metrics to explicitly consider diversity (view *section 3.3.2*). Finally, through an empirical evaluation, the

authors demonstrate that *IA-Select* scores higher on their proposed metrics compared to commercial search engines.

In the first place, the existence of a taxonomy of information is introduced, where both documents \mathbb{D} and queries q can belong to various categories c from the defined taxonomy. In addition, information on overall user intents is known and complete. This information is represented by the probability distribution $P(c|q)$, which defines the probability that the given query q belongs to category c . The relevance/quality of a document d in response to a query q with the intended category c , is captured by the probability of $V(d|q, c)$. Given the probability distribution $P(c|q)$ and the relevance values $V(d|q, c)$, the problem of result diversification is to find a set of documents $\mathbb{S} \subseteq \mathbb{D}$ of size k that maximizes *Equation 10*.

$$Diversify(\mathbb{D}, \mathbb{S}, k, q) \stackrel{\text{def}}{=} \arg \max_{\mathbb{S} \subseteq \mathbb{D} : |\mathbb{S}|=k} \left(\sum_c \left(P(c|q) \cdot \left(1 - \prod_{d \in \mathbb{S}} (1 - V(d|q, c)) \right) \right) \right) \quad (10)$$

Equation 10. Agrawal diversification optimization function [Agra09]

Where, $\prod_{d \in \mathbb{S}} (1 - V(d|q, c))$ is the probability that the user will find none of the documents in \mathbb{S} relevant to the query q and intent c . Thus $(1 - \prod_{d \in \mathbb{S}} (1 - V(d|q, c)))$, indicates the probability that the user will find in \mathbb{S} at least one relevant document to satisfy q when the intent is c . As can be seen the overall goal is for the user to find at least one relevant document, and in consequence, minimize the risk of dissatisfying the average user. Note that this optimization goal implicitly promotes diversity. If a highly relevant document from a category c has already been added to \mathbb{S} , then the gain of adding another document also belonging to c is very small.

Agrawal *et al.* [Agra09] explain that their optimization objective is *NP-hard*. However, the authors demonstrate the submodularity of their function and conclude that it can be solved with a greedy strategy. This strategy is *IA-Select*, which is an algorithm able to find an optimal solution with bounded error to *Equation 10*.

IA-Select receives as input documents retrieved from a classical ranking algorithm in response to the query. At each step, *IA-Select* chooses a document to add to \mathbb{S} that has the highest marginal gain, defined in *Equation 11*.

$$g(d|q, c, \mathbb{S}) = \sum_{c \in C(d)} U(c|q, \mathbb{S}) V(d|q, c) \quad (11)$$

Equation 11. Marginal gain in IA-Select algorithm [Agra09]

Where $C(d)$ are the categories that document d belongs to, and $U(c|q, \mathbb{S})$ is the probability that q belongs to c if all documents in \mathbb{S} fail to satisfy the query q . Agrawal *et al.* [Agra09] explain: “this marginal utility can be interpreted as the probability that the selected document satisfies the user given that all documents that come before it fail to do so”. When \mathbb{S} is empty, $U(c|q, \emptyset) = P(c|q)$. When a document d^* is chosen at a step, for all categories d^* belongs to their corresponding $U(c|q, \mathbb{S})$ are updated as defined in *Equation 12*. As can be seen, $U(c|q, \mathbb{S})$ decreases in proportion to the relevance/quality of the selected document d^* towards satisfying q with intent c .

$$U(c|q, \mathbb{S}) = (1 - V(d^*|q, c)) \cdot U(c|q, \mathbb{S} \setminus \{d^*\}) \quad (12)$$

Equation 12. Update of $U(c|q, \mathbb{S})$ in IA-Select algorithm [Agra09]

3.1.1.3 The *xQuAD* Framework [Sant10]

Santos *et al.* [Sant10] propose the probabilistic framework *xQuAD* (eXplicit **Q**uery Aspect **D**iversification) for web search result diversification. This framework considers both relevance and diversity to re-rank documents in such a way that maximum coverage over different sub-topics/sub-queries/aspects of the query is obtained while simultaneously minimizing redundancy. In addition, Santos *et al.* use query reformulations obtained by existing search engines in order to further uncover query aspects. The authors carried out offline experiments using the guidelines and dataset provided by the TREC 2009 diversity task. They found that in various settings their approach outperforms several state-of-the-art techniques.

The *xQuAD* framework uses a greedy approach to iteratively select a set \mathbb{S} of documents from a broader set \mathbb{R} that maximizes *Equation 13*.

$$xQuAD(\mathbb{Q}, \mathbb{R}, \mathbb{S}, \lambda) \stackrel{\text{def}}{=} \arg \max_{d \in \mathbb{R} \setminus \mathbb{S}} \left((1 - \lambda) \cdot P(d|\mathbb{Q}) + \lambda \cdot P(d, \bar{\mathbb{S}}|\mathbb{Q}) \right) \quad (13)$$

Equation 13. xQuAD framework objective function [Sant10]

In *Equation 13*, \mathbb{Q} is the ambiguous query and \mathbb{R} is an initial ranking generated for the query from a traditional *IR* system. The probability model mixes: (i) *Relevance*: $P(d|\mathbb{Q})$ which is the likelihood document d is observed given \mathbb{Q} , and (ii) *Diversity*: $P(d, \bar{\mathbb{S}}|\mathbb{Q})$ which is the probability of observing d but not documents that are already in \mathbb{S} given \mathbb{Q} . The parameter λ controls the trade-off between relevance and diversity. The probability $P(d, \bar{\mathbb{S}}|\mathbb{Q})$ is calculated by explicitly considering the different query aspects q_i as in *Equation 14*.

$$P(d, \bar{\mathbb{S}}|\mathbb{Q}) = \sum_{q_i \in \mathbb{Q}} P(q_i|\mathbb{Q}) \cdot P(d, \bar{\mathbb{S}}|q_i) \quad (14a)$$

$$P(d, \bar{\mathbb{S}}|q_i) = P(d|q_i) \cdot P(\bar{\mathbb{S}}|q_i) \quad (14b)$$

$$P(\bar{\mathbb{S}}|q_i) = \prod_{d_j \in \mathbb{S}} (1 - P(d_j|q_i)) \quad (14c)$$

Equation 14. xQuAD diversity probability [Sant10]

In this manner *Equation 13* is rewritten as *Equation 15*.

$$xQuAD(\mathbb{Q}, \mathbb{R}, \mathbb{S}, \lambda) \stackrel{\text{def}}{=} \arg \max_{d \in \mathbb{R} \setminus \mathbb{S}} \left((1 - \lambda) \cdot P(d|\mathbb{Q}) + \lambda \cdot \sum_{q_i \in \mathbb{Q}} \left(P(q_i|\mathbb{Q}) \cdot P(d|q_i) \cdot \prod_{d_j \in \mathbb{S}} (1 - P(d_j|q_i)) \right) \right) \quad (15)$$

Equation 15. xQuAD framework objective function rewritten [Sant10]

Santos *et al.* in [Sant10a] continue their work and propose a way to optimally learn λ depending on the given query. They explain that different queries would benefit from different diversity-relevance trade-offs given that they have different levels of ambiguity. In this manner, for an unseen query Santos *et al.* predict λ based on results of similar previously seen queries. Finally, with offline experiments, the authors demonstrate that their approach for selectively choosing λ outperforms having a uniform diversification strategy.

3.1.1.4 Coverage-based search result diversification [Zhen12]

In [Zhen12], *IR* diversification is viewed as an optimization problem that aims to maximize a coverage-based diversity function. The goal is to obtain a ranked list of documents that covers

as many different query sub-topics as possible. The authors propose three different strategies to generate coverage functions that are based on summations, loss functions and evaluation measures. From these strategies different coverage functions are obtained. Some of the functions are corresponding to several state-of-the-art methods and others result in completely new diversification methods. The authors compare the methods both analytically and empirically. They show that all methods are effective and one of the new methods even outperforms existing ones.

In first place, Zheng *et al.* [Zhen12] establish a framework that can accommodate the different coverage functions. In their framework, the authors propose a greedy strategy that aims at optimizing a submodular objective function given by *Equation 16*. This function takes into account the trade-off between diversity and relevance. The diversity of a document d , is measured considering the importance/weight and number of sub-topics associated to the query q that the document d covers within the context of previously selected documents \mathbb{R} . In *Equation 16*, $S(q)$ represents the sub-topics related to query q . In other words, the more sub-topics of q that a document covers, the more important the covered topics are, and the more dissimilar these topics are from previously selected topics; the higher the document's diversity is. The challenge now is to determine possible coverage functions $cov(s, d, \mathbb{R})$ and to prove that they are submodular.

$$d^* = \arg \max_{d \in \mathbb{D} \setminus \mathbb{R}} \left(\lambda \cdot rel(q, d) + (1 - \lambda) \cdot \sum_{s \in S(q)} (weight(s, q) \cdot cov(s, d, \mathbb{R})) \right) \quad (16)$$

Equation 16. Coverage-based framework for diversification in IR [Zhen12]

The authors highlight five different structures for coverage functions. These structures are designed to decrease the gain of adding a document covering a sub-topic that has already been well covered. From the generated functions the authors find they can represent related works such as *IA-Select* [Agra09] (view *section 3.1.1.2*) and *xQuAD* [Sant10] (view *section 3.1.1.3*). Furthermore, they find that the coverage function structure in *Equation 17* outperforms related works by conducting experiments with all methods over two TREC dataset.

$$cov(s, d, \mathbb{R}) = cov(s, d) \cdot \left(2 - 2 \sum_{d' \in \mathbb{R}} cov(s, d') - cov(s, d) \right) \quad (17)$$

Equation 17. Squared loss coverage function [Zhen12]

In order to complete the framework what is left is to define how to measure $cov(s, d)$ and $weight(s, q)$. In their work, Zheng *et al.* [Zhen12] define these functions as in *Equation 18*, where $P(s|q)$ is the probability that sub-topic s is relevant to query q , and $P(d|s)$ is the probability that document d is relevant given sub-topic s . This is also inspired on the previous works of Agrawal *et al.* [Agra09] (view *section 3.1.1.2*) and Santos *et al.* [Sant10] (view *section 3.1.1.3*).

$$weight(s, q) = P(s|q) \quad (18a)$$

$$cov(s, d) = P(d|s) \quad (18b)$$

Equation 18. Defining coverage and weight functions [Zhen12]

In this section we have presented four *IR* diversification techniques. In the following section we will *RecSys* diversification techniques will be described. To conclude, a discussion comparing all presented techniques, both for *IR* and *RecSys*, will be offered in *section 3.2*.

3.1.2 Diversification Techniques for Recommendation Systems

In related works, two lines of research can be identified that propose solutions to the diversification problem in Recommendation Systems. The first and strongest line focuses on post-filtering approaches. These approaches receive as input recommendation results generated by a traditional *RecSys* algorithm, and aim to select from the *candidate items* the best subset that balances diversity and quality/accuracy to generate a final item recommendation list (view *Figure 5*). The second line of proposed solutions attempts to enhance current *RecSys* algorithms in order to generate more diverse item recommendations.

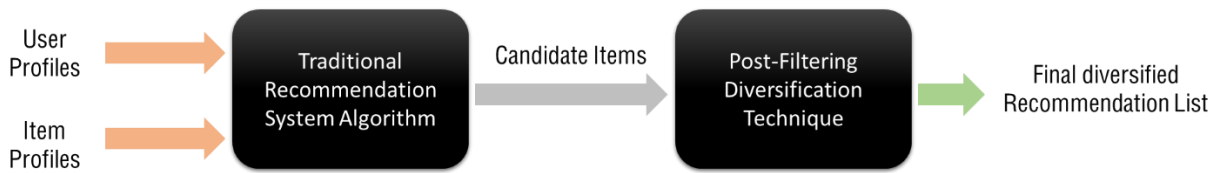


Figure 5. Post-filtering Diversification Techniques

It is important to clarify that diversity can be achieved at many levels. This research focuses on individual intra-list diversity, *i.e.*, offering one user one diversified list of item recommendations. Other types of diversity are focused on individual inter-list diversity and on aggregate diversity. On the one hand, individual inter-list diversity is concerned with offering one user lists of recommended items that are diverse within themselves over time. Lathia *et al.* [Lath10] study the temporal characteristics of diversity in *RecSys* emphasizing that there is no current mechanism to prevent users from receiving the same recommendations over and over again. On the other hand, Adomavicius and Kwon [Adom12Adom12] introduce the concept of aggregate diversity as offering diverse recommendations across all users. They highlight that aggregate diversity is not necessarily achieved through individual diversity, *e.g.*, if all users get the same diverse list of recommendations the result is high individual diversity but low aggregate diversity.

In this section both post-filtering and *RecSys* enhancing diversification methods are discussed for the goal of individual intra-list diversity.

3.1.2.1 Topic Diversification [Zieg05]

Ziegler *et al.* in [Zieg05] present the *Topic Diversification* method, which is intended to “*balance and diversify personalized recommendations lists in order to reflect the user’s complete spectrum of interests*”. In addition, they propose the *Intra-List Similarity* metric which is further discussed in *section 3.3.4*. Ziegler *et al.*, evaluate their method using book recommendation data with both offline and online studies. Offline studies intended to understand the impact of Topic Diversification on accuracy metrics. Online studies intended to understand how their method affected real user satisfaction. The authors found that even though increasing diversity is detrimental to accuracy it improves real user satisfaction.

The Topic Diversification Algorithm considers as input a recommendation list \mathbb{P}_{w_i} generated by an existing recommender system. The algorithm’s goal is to re-rank the input list \mathbb{P}_{w_i} and

obtain a diversified final top-N list $\mathbb{P}_{w_i^*}$. The input list \mathbb{P}_{w_i} must be considerably larger than the desired list $\mathbb{P}_{w_i^*}$. The algorithm uses a diversification factor Θ_F to control the trade-off between relevance and diversity. In addition, a content-based similarity measure is used to compare items. The Topic Diversification (*TopicDiv*) algorithm attempts to greedily select items from \mathbb{P}_{w_i} to add to $\mathbb{P}_{w_i^*}$ which minimize the function in *Equation 19*.

$$TopicDiv(\mathbb{P}_{w_i}, \mathbb{P}_{w_i^*}, \Theta_F) \stackrel{\text{def}}{=} \min_{b \in \mathbb{P}_{w_i} \setminus \mathbb{P}_{w_i^*}} \left((1 - \Theta_F) * rank(\mathbb{P}_{w_i}(b)) + \Theta_F * rank(\mathbb{P}_{c^*}^{rev}(b)) \right) \quad (19)$$

Equation 19. Topic Diversification [Zieg05]

Where $\mathbb{P}_{w_i}/\mathbb{P}_{w_i^*}$ is the set difference between \mathbb{P}_{w_i} and $\mathbb{P}_{w_i^*}$; and $\mathbb{P}_{c^*}^{rev}(b)$ is called the dissimilarity rank. The dissimilarity rank a list sorted defined as *Equation 20*.

$$\mathbb{P}_{c^*}^{rev}(b, \mathbb{P}_{w_i^*}) = \arg \min_{p \in \mathbb{P}_{w_i^*}} sim(b, p) \quad (20)$$

Equation 20. Dissimilarity rank for Topic Diversification [Zieg05]

The goal is to minimize the weighted linear combination of two rank values (*i.e.*, position in list): (a) $rank(\mathbb{P}_{w_i}(b))$ returns the position in the ordered list \mathbb{P}_{w_i} of item b (\mathbb{P}_{w_i} is ordered by the predicted rating for items), the smaller the value of the rank position the higher the relevance; and (b) $rank(\mathbb{P}_{c^*}^{rev}(b))$ is the position of item b in the dissimilarity rank, the smaller the rank position the more dissimilar b is to items already in $\mathbb{P}_{w_i^*}$.

The authors make an analogy to Osmotic Pressure, and emphasize that Topic Diversification resembles the membrane's selective permeability known from molecular biology. Vargas *et al.* [Varg12] point out that Ziegler's *et al.* approach is very similar to the *MMR* method proposed in by Carbonell and Goldstein [Carb98] (view *section 3.1.1.1*).

3.1.2.2 Similarity vs. Diversity in Case-based Recommendation Systems [Smyt01]

Smyth and McClave [Smyt01] argue that, even though similarity-based retrieval strategies are usually preferred in the field of case-based reasoning, often diversification could be equally as important; especially in Case-Based Recommender Systems (*CBR*). They explain: “Often, it is not good enough for a recommender to return only the most similar cases. It should also return a diverse set of cases in order to provide the user with optimal coverage of the information space in the vicinity of their query” [Smyt01]. As a solution, the authors propose and evaluate strategies to diversify results from *CBR* that do not compromise similarity or performance.

In their work, the authors define dissimilarity as the inverse of similarity as in *Equation 6*, and define diversity as pairwise dissimilarity as in *Equation 2*. With this foundation, the authors propose three algorithms viewed in *Figure 6*: *BoundedRandomSelection*, *GreedySelection* and *BoundedGreedySelection*. In first place, the *BoundedRandomSelection* strategy identifies the b_k cases most similar to the target query and chooses from those at random k cases, where $k \leq b_k$, to define the result set. In second place, the *GreedySelection* chooses at each step the case that maximizes a quality criterion $Quality(t, c, R)$ to return a result list of k cases. Lastly, the *BoundedGreedySelection* is an adapted version of the *GreedySelection* with a bounded input restricted to the b_k most similar cases to the target query. The authors show that the

GreedySelection algorithm is expensive in performance and that the *BoundedGreedySelection* is a faster alternative that produces comparable high quality results.

<pre> t: target query, C: case-base, k: # results, b: bound 1. define BoundedRandomSelection (t, C, k, b) 2. begin 3. C' := bk cases in C that are most similar to t 3. R := k random cases from C' 4. return R 5. end </pre>
<pre> 1. define GreedySelection (t, C, k) 2. begin 3. R := {} 4. For i := 1 to k 5. Sort C by Quality(t,c,R) for each c in C 6. R := R + First(C) 7. C := C - First(C) 8. EndFor 9. return R 10. end </pre>
<pre> 1. define BoundedGreedySelection (t, C, k, b) 2. begin 3. C' := bk cases in C that are most similar to t 4. R := {} 5. For i := 1 to k 6. Sort C' by Quality(t,c,R) for each c in C' 7. R := R + First(C') 8. C' := C' - First(C') 9. EndFor 10. return R 11. end </pre>

Figure 6 Diversity preserving algorithms. Figure extracted from [Smyt01].

The authors propose the quality metrics defined in Equation 21 for the greedy selection algorithms. In these metrics, t is the target query, c is the case base being evaluated and R is the set of cases that have already been selected by the algorithm. As can be seen Equation 21b is very similar to the objective function proposed as *MMR* (view section 3.1.1.1).

$$Quality(t, c, R) = Similarity(t, c) \cdot RelDiversity(c, R) \quad (21a)$$

$$Quality(t, c, R) = \alpha \cdot Similarity(t, c) + (1 - \alpha) \cdot RelDiversity(c, R) \quad (21b)$$

$$Quality(t, c, R) = \frac{2}{\left(\frac{1}{Similarity(t, c)} + \frac{1}{RelDiversity(c, R)} \right)} \quad (21c)$$

Equation 21. Quality metrics in [Smyt01]

3.1.2.3 Item Re-Ranking Methods [Adom09]

Adomavicius and Kwon [Adom09] propose item re-ranking methods to increase aggregate diversity while maintaining acceptable levels of accuracy for recommendation results. Aggregate diversity is defined as the diversity obtained from recommendations across all users, measured as the “total number of distinct items recommended across all users” [Adom09].

Adomavicius and Kwon explain that traditional recommender systems offer to a user a list of top- N item recommendations ordered according to a ranking criterion. The standard ranking criterion sorts items in decreasing order according to the predicted relevance/rating of each item. However, this standard ranking is designed to improve accuracy but does not consider

diversity. This is why the authors propose new ranking methods that can at the same time control accuracy losses.

In the first place, to control the accuracy-diversity trade-off the authors propose a ranking function $rank_X(i, T_R)$ parameterized by a *ranking threshold* T_R . The ranking threshold must comply with $T_R \in [T_H, T_{max}]$, where T_{max} is the highest possible rating and T_H is a threshold for *high* ratings. The predicted rating for an item is defined by $R^*(u, i)$ and is considered highly-ranked or relevant if $R^*(u, i) \geq T_H$. The proposed ranking function is then defined as in *Equation 22*.

$$rank_X(i, T_R) = \begin{cases} rank_X(i), & \text{if } R^*(u, i) \in [T_R, T_{max}] \\ \alpha_u + rank_{standard}(i), & \text{if } R^*(u, i) \in [T_H, T_R] \end{cases} \quad (22)$$

Equation 22. Re-ranking Method proposed by [Adom09]

Where α_u is defined as in *Equation 23*.

$$\alpha_u = \max_{i \in (R^*(u, i) \geq T_R)} rank_X(i) \quad (23)$$

Equation 23. α_u in Re-ranking Method proposed by [Adom09]

In this manner, items that have predicted rating $R^*(u, i) \geq T_R$ are ranked with $rank_X(i)$ and items with $R^*(u, i) < T_R$ are ranked with the standard ranking $rank_{standard}(i)$. Also, α_u ensures that all items with $R^*(u, i) \geq T_R$ get ranked ahead of all items with $R^*(u, i) < T_R$. As can be seen T_R controls how much of the original accuracy is kept in the new rank.

The authors propose five ranking functions for $rank_X(i)$ shown in *Figure 7*.

<ul style="list-style-type: none"> • Standard, i.e., ranking the candidate (highly predicted) items by their predicted rating value, from highest to lowest. $rank_{Standard}(i) = R^*(u, i)^{-1}$. • Item Popularity, i.e., ranking by item popularity, from lowest to highest, where popularity is represented by the number of known ratings that each item has. $rank_{ItemPop}(i) = U(i)$, $U(i) = \{u \in U \mid \exists R(u, i)\}$. • Reverse Predicted Rating Value, i.e., ranking by predicted rating value, from lowest to highest. $rank_{RevPred}(i) = R^*(u, i)$. 	<ul style="list-style-type: none"> • Item Average Rating, i.e., ranking by an average of all known ratings for each item: $rank_{AvgRating}(i) = \overline{R}(i) = \sum_{u \in U(i)} R(u, i) / U(i)$. • Item Absolute Likeability, i.e., ranking by how many users liked the item (i.e., rated it above T_H): $rank_{AbsLike}(i) = U_H(i)$, $U_H(i) = \{u \in U(i) \mid R(u, i) \geq T_H\}$. • Item Relative Likeability, i.e., ranking by the percentage of the users who liked an item (among all users who rated it). $rank_{RelLike}(i) = U_H(i) / U(i)$.
--	--

Figure 7. Re-ranking strategies proposed by [Adom09]

The authors carried out offline tests with two datasets: MovieLens and Netflix. The five ranking approaches were tested with input from three collaborative filtering techniques: user-based, item-based and matrix factorization. Each ranking was measured in terms of *precision-in-top-N* and *diversity-in-top-N* (aggregate diversity). Also the diversity gain and precision loss with respect to the standard ranking was measured. They found that all ranking approaches sacrificed accuracy to gain diversity. However, different values for the ranking threshold T_R allowed control over the diversity-accuracy trade-off. The key is then to find a threshold that offers high diversity gain at tolerable precision loss. An interesting effect discovered by the authors is that an increase in diversity also leads to an increase in the number of long-tail items recommended.

3.1.2.4 Information Retrieval Diversity for Recommender Systems [Varg12]

Vargas [Varg12] identified an opportunity in adapting diversity metrics and techniques from *IR* to *RecSys* given that recommendation can be viewed as an *IR* task. He argues that diversity in *IR* has better conceptual foundations and a “*drive towards standardization (backed by a specific TREC diversity task)*” that is not yet present in *RecSys*. In this manner, it is natural to attempt to transfer knowledge from one area to the other and thus benefit research on *RecSys* and vice-versa.

To achieve this, Vargas [Varg12] proposes the concept of *aspect space*. The aspect space serves as a mean to translate equivalent notions from *IR* to *RecSys*, specifically: document similarity from *IR* is corresponding to item similarity from *RecSys* and query intents from *IR* is analogous to user profile aspects in *RecSys*. With this analogy, techniques and metrics associated to diversity from *IR* can be easily adjusted towards *RecSys*.

An *aspect* represents a disjoint user interest found within the user profile. As user interests are not equally important, Vargas defines the aspect space as a probability distribution retrieved from the user profile defined as $p(a|u)$. Similar to query ambiguity, there is uncertainty associated to the intent that would satisfy the user the most in a given context.

Correspondingly, there must be a way to determine how aspects are covered by items. Vargas considers for each item an aspect space defined by $p(a|i)$. Also, with item aspect spaces it is possible to derive similarity metrics to compare items.

To show the application of the aspect space, Vargas adapts the *IA-Select* algorithm (view section 3.1.1.2) to diversify results in *RecSys*. In first place, Vargas rephrases the gain function from *IA-Select* (seen in Equation 24a) to incorporate the updates of $U(c|q, \mathbb{S})$ as can be viewed in Equation 24c. To define the aspect space, the taxonomy of categories c are viewed as aspects a . The adaptation of *IA-Select* can be viewed in Equation 25.

$$g(d|q, c, \mathbb{S}) = \sum_{c \in C(d)} U(c|q, \mathbb{S})V(d|q, c) \quad (24a)$$

$$U(c|q, \mathbb{S}) = p(c|q) \cdot \prod_{d' \in \mathbb{S}} (1 - V(d'|q, c)) \quad (24b)$$

$$g(d|q, c, \mathbb{S}) = \sum_{c \in C(d)} p(c|q) \cdot V(d|q, c) \prod_{d' \in \mathbb{S}} (1 - V(d'|q, c)) \quad (24c)$$

Equation 24. Modified Marginal gain from *IA-Select* [Agra09][Varg12]

$$g(i|u, a, \mathbb{S}) = \sum_{a \in \mathcal{A}} \left(p(a|u) \cdot \hat{r}_{norm}(u, i) \cdot p(a|i) \cdot \prod_{j \in \mathbb{S}} (1 - \hat{r}_{norm}(u, j) \cdot p(a|j)) \right) \quad (25a)$$

$$\text{Where, } V(i|u, a) = \hat{r}_{norm}(u, i) \cdot p(a|i) \quad (25b)$$

Equation 25. Adaptation of *IA-Select* using the aspect space [Agra09][Varg12]

Vargas explains that when information about item features and the user profile can be accessed then the aspect space is explicitly defined. In this case, features f are considered equivalent to aspects and the aspect space is defined as in Equation 26.

$$p(a|i) = p(f|i) = \frac{|f \in i|}{|i|} \quad (26a)$$

$$p(a|u) = p(f|u) = \frac{[i \in u | f \in i]}{\sum_{f' \in \mathcal{F}} [i \in u | f' \in i]} \quad (26b)$$

Equation 26. Modified Marginal gain from IA-Select [Agra09][Varg12]

However, item feature information is sometimes missing or incomplete. In this case an implicit aspect space must be defined from the user profile. Vargas proposes the use of matrix factorization to extract latent features from the user profile to be defined as aspects.

Vargas tests the behavior of the adapted diversification algorithms using the MovieLens 100k dataset. As a baseline, Vargas takes two collaborative filtering (*CF*) algorithms: user-user *CF* and matrix factorization. The diversification algorithms re-rank the top 500 items returned by the baseline *RecSys*. Vargas carries out tests for both implicit and explicit aspect space scenarios. He shows that the adapted diversification algorithms perform better than the non-diversified baselines using as error metrics α -*nDCG@50*, *ERR-IA@50*, *nDCG-IA@50* and *ILD@50* (view [Varg12] and *section 3.3* for more information). Also, Vargas found that implicit aspect space performs better in many cases than explicit aspect space and this could be explained by the fact that latent features tend to be a “*more dense representation of items*” [Varg12].

3.1.2.5 Latent Factor Portfolio [Shi12]

Shi *et al.* [Shi12], inspired on previous work on applying *Modern Portfolio Theory (MPT)* to *IR* [Wang09], propose the Latent Factor Portfolio method to diversify results produced from a collaborative filtering *RecSys* based on a latent factor model (*e.g.*, matrix factorization models such as singular value decomposition and *SVD++*). Their proposal is different because the diversification level of results is adapted to the uncertainty of the user profile.

In [Wang09] the authors explain that the ranking task in *IR* resembles an investment problem where the *IR* system is *investing on raking positions in documents*. They present *MPT* as a possible approximation to find an optimal portfolio/ranking. In *MPT* the following observations are made: the future return on stock is unknown, investors have different preferences of risk, and future return of stocks can be correlated and therefore choosing stock independently to build a portfolio is undesirable. *MPT* aims to select a portfolio that maximizes *expected returns* and minimizes *risk*. The challenge lies on the observation that a stock with higher returns is generally riskier. Also, it must be considered that adding too many stocks that are positively correlated increases the risk of the portfolio (*i.e.*, if one goes down then the others do as well). In general, the optimal portfolio is a portfolio composed of diverse stock investments.

Shi *et al.* [Shi12] view a clear connection between *MPT* and diversification in *RecSys*. Like stock, the expected return of items (*i.e.* the rating for an unrated item) is unknown, and cannot be estimated with absolute certainty from the user profile. Also, the expected return of items is correlated either positively or negatively. Shi *et al.* [Shi12] propose the Latent Factor Portfolio method to diversify recommendation results. In their method the expected return can be measured as the expected value of a rating and the risk is associated to the uncertainty of the user profile measured by the variance in the latent factors within the profile.

Specifically, the Latent Factor Portfolio method ranks an item at a given position based not only on its predicted rating, but also based on its uncertainty in terms of the latent item factors and the correlation between the item and the items ranked before it. On the one hand, the uncertainty of the latent item factors are obtained from the uncertainty of the same latent factors

found in the user profile (*i.e.*, the variances of the latent factors in the user profile). On the other hand, the correlation between items (diversity) is obtained from the correlation between their latent factors.

The authors establish that by considering the uncertainty of latent factors in the user profile the recommendation diversity is adjusted to respond to the user's demand on diversity. As a result, their method produces more diverse recommendations for users with diverse user profiles, assuming that if the user has specific interests then less diversity is required. Even though the authors highlight this adaptive quality as an important part of their contributions, we think it's actually detrimental to diversity. A user with a homogenous user profile in given situations could benefit from a diversified list of results. Also, an important observation is that the Latent Factor Portfolio penalizes novel items. In a latent factor based recommender, novel items will tend to have lower predicted ratings compared to other items that are more related to the user profile. In this setting, novel items are characterized as items with high risk and low expected return and would have difficulty being recommended.

3.1.2.6 User Profile Partitioning [Zhan09a][Varg13]

In user profile partitioning diversification techniques, items in the user profile are grouped in order to identify the user's underlying preferences. Next, recommended items are selected in order to cover as many of the user preferences as possible. This section reviews two solutions based on a user profile partitioning approach. Both follow the subsequent three basic steps: (i) partition/cluster the user profile, (ii) generate recommendations for each of the "*sub-profiles*" by treating each partition as if it were an independent user profile, and (iii) aggregate recommendations.

Zhang and Hurley [Zhan09a] propose a new collaborative-filtering recommendation algorithm that increases the probability of recommending items that are both novel and relevant. Novel items are defined as those that correspond to more unusual tastes of the user and that the recommender finds hard to suggest. The authors explain that it is unproductive to diversify the results of traditional recommendation systems if it is known that generated recommendations tend to be biased towards non-novel items. In addition, they explain that: "*novel recommendation is difficult because standard similarity metrics measure the aggregate similarity to multiple items in the user profile and the influence of more novel items is lost in the aggregation*" [Zhan09a]. In order to obtain more novel items, the authors suggest it is better to offer recommendations from the individual clusters within the user profile and not the user profile as whole. Their goal is to increase the probability of recommending novel items that can cover the user's range of tastes. To do so, they explore a number of user profile partitioning techniques such as: maximization of the intra-cluster similarity, graph partitioning, *k*-means, and others. After identifying the user preference clusters, these are sorted according to the average novelty of the items they contain. Only *k* clusters with the largest aggregate novelty are used for making recommendations. The authors propose a new evaluation methodology that can capture the *RecSys*'s ability to offer diversified results across relevant items independent of their novelty. They evaluate their algorithm on the MovieLens dataset to show that their algorithm succeeds at removing bias towards non-novel items at low accuracy cost.

It is important to highlight one of the conclusions of Zhang and Hurley's work: "*While achieving significant improvement over standard algorithms, our best performing algorithms still obtain high Gini index values, showing that hits are still concentrated among a relatively*

small set of items” [Zhan09a]. The authors explain that in future works they will address new ways to enhance diversification. However, this is an interesting conclusion because it shows that even if generated recommendations have novel items, recommending items with the standard relevance-maximization ranking is not enough to obtain diversity.

On a similar work, Vargas and Castells [Varg13] partition the user profile into sub-profiles using known information about categories in the item domain. Next, they generate item recommendations for each sub-profile. In order to generate recommendations for sub-profiles with collaborative filtering methods, the authors analyze different possibilities such as: should sub-profile similarity be calculated only with other user sub-profiles or with complete user profiles?. After recommendations are generated, the final challenge is to combine results and obtain a final diversity-aware recommendation list. To achieve this, the authors use an adjusted version of the *xQuAD* diversification algorithm (view *section 3.1.1.3*). The original *xQuAD* diversification algorithm for *IR* is first modified to be used for *RecSys* based on the idea of aspect spaces as explained in *section 3.1.2.4*. Next, the diversification method is further adapted to consider user sub-profiles and not the complete user profile. The authors carry out tests with the MovieLens 1M dataset and Last.fm dataset showing their proposal achieves competitive results.

3.1.2.7 Diversifying User Neighbors [Smyt01][Yang13][Zhan09][Said12]

Smyth *et al.* [Smyt01], as possible application for their diversification techniques (view *section 3.1.2.2*) and as future work, propose that diversification can be used by collaborative filtering algorithms to retrieve k diverse similar neighbors instead of the k most similar neighbors to the target user. As can happen when only using similarity metrics, a set of user profiles that are the most similar to a target user can also be very similar to each other, and therefore contain item ratings for the same small set of items. This limits the possible recommendation space. If instead of selecting the most similar users a diverse set of users is chosen as the neighborhood, then a larger space of items can be evaluated to therefore generate a more diverse recommendation list. The authors also point out that it is not enough to change the similarity metric to achieve diversity. They state: “*Similarity and diversity are orthogonal measures. Similarity is a local function of two cases, the target and a candidate, and the similarity of a case with respect to a target does not depend on the similarity of any other case. In contrast, the relative diversity of a case depends on previous similarity computations (and case selections). For this reason it is not possible to fold diversity in to a single similarity computation*” [Smyt01].

Along the same motivation, Yang *et al.* [Yang13] propose a neighbor diversification-based collaborative filtering algorithm that selects neighbors by means of a Neighbor Diversification Algorithm. Recommendations are then generated for the diverse set of neighbors. By experimentation with the MovieLens dataset, the authors prove that through diverse neighbor selection recommendation diversity, novelty and coverage can be improved keeping and sometimes even improving accuracy.

Similarly, Zhang [Zhan09] proposes a recommendation algorithm that selects diverse neighbors in a trust-based *RecSys*. Zhang argues that the limitation of post-filtering diversification techniques is that the input candidate items (*i.e.*, the recommendations generated from a traditional *RecSys* technique) might not be diverse enough, and as a consequence, the best subset that can be selected by a post-filtering algorithm will not be diverse enough as well. In

these cases, the *RecSys* algorithm itself must be enhanced to generate diverse results. The author proposes a greedy optimization diversification strategy in order to select a diverse set of user neighbors in a trust-based collaborative filtering algorithm. The objective function proposed has the same structure as *MMR* (view *Equation 9*), where the relevance of a neighbor is interpreted as the trust value and the diversity is measured with traditional user similarity metrics.

As a different approach, Said *et al.* [Said12] propose a *k*-furthest neighbors collaborative filtering technique to increase serendipity and diversity. This technique recommends items that are disliked by users which are the least similar to the target user. Through experiments with the MovieLens dataset, the authors find that their method provides higher diversity at a tolerable precision loss compared to the traditional nearest neighborhood technique.

In this section *RecSys* diversification techniques have been presented. In the following section a comparative analysis of techniques for both *IR* and *RecSys* will be offered.

3.2 Discussion on Diversification Techniques

In this section we have presented approaches that follow two lines of research:

- (a) *Post-filtering Approaches*: take as input results generated from a traditional *RecSys* and select a subset of diversified quality items.
- (b) *Enhancing Traditional Recommendation Algorithms Approaches*: improve traditional *RecSys* algorithms in order to incorporate diversity-awareness.

In *Figure 8*, reviewed works are classified between the two lines of research.

Post-filtering approaches for *RecSys* tend to draw the most inspiration from previous *IR* diversification approaches. On the one hand, in Ziegler's *et al.* [Zieg05] and Smyth's *et al.* [Smyt01] work, inspiration can be seen from *MMR* [Carb98]. On the other hand, Vargas [Varg12] proposes a method to adapt *IR* diversification techniques, such as *xQuAD* [Sant10] and *IA-Select* [Agra09], to the field of *RecSys*.

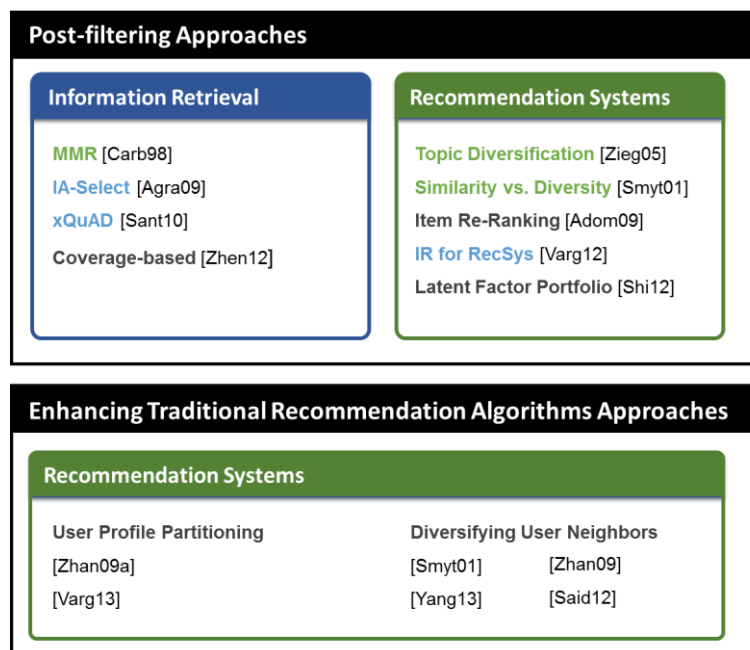


Figure 8. Diversification Techniques

Proposed solutions from both lines of research can be easily used together to create diversity-aware *RecSys*: the output from the diversification enhanced recommendation algorithm could serve as input for the post-filtering diversification approach. Even though both approaches benefit each other, the only example of a combined solution that we know of is proposed by Vargas in [Varg13] (view *section 3.1.2.7*). A combined approach seems to be an ideal approximation: (i) using a post-filtering approach alone if the set of candidate items is not diverse enough could generate a sub-optimal result as pointed out by [Zhan09] (view *section 3.1.2.7*), and (ii) enhancing a traditional *RecSys* algorithm to generate diverse recommendations but ranking these by relevance could also produce sub-optimal results as found by [Zhan09a] (view *section 3.1.2.6*). Regrettably, given the scope of this research we must concentrate in one of the two lines of research. We will concentrate on post-filtering approaches in order to enhance collaborative filtering algorithms which compared to content-based techniques produce more novel and diverse results. In the future, the proposed post-filtering diversification approach could be adapted to enhance a *RecSys* algorithm. However, combined approaches will be researched in future works.

From this point forward we will analyze post-filtering approaches. To begin with, criteria used in order to compare the studied approaches is defined in *Table 3*.

Criteria	Definition
<i>Greedy Optimization</i>	Is the proposed solution a greedy optimization approach?
<i>Explicit Approach</i>	Does the proposed solution directly attempt to cover the diverse aspects of the query/user profile?
<i>Implicit Approach</i>	Does the proposed solution explicitly prevent redundancy within the results?
<i>Control of diversity vs. relevance trade-off</i>	Is there a control parameter that can tune the diversity vs. relevance trade-off?
<i>Encourages Discovery</i>	Does the proposed approach not penalize novel/serendipitous items?
<i>Control of exploitation vs. exploration trade-off</i>	Is there a control parameter that can tune the exploitation vs. exploration trade-off?
<i>Considers overlapping item categories</i>	Does the proposed approach consider that item characteristics could be interdependent?

Table 3. Comparison Criteria for Diversification Techniques

Next, a detailed clarification of criteria is presented:

(a) *Greedy Optimization*

This criterion serves to highlight approaches that are based on a greedy optimization solution. It must be clarified that this is not a required, or necessary desirable, characteristic for diversification techniques. Even so, knowing if a diversification technique is based on a greedy approach offers important insight to compare techniques.

(b) *Explicit vs. Implicit Approaches*

According to [GilC13][Varg13][Zhen12][Sant10] diversification approaches for both *RecSys* and *IR* can be classified as *Implicit* or *Explicit*. In *IR*, *implicit* approaches infer that by selecting

dissimilar items the diverse query aspects will be indirectly covered. *MMR* [Carb98] is an example of an implicit approach. In contrast, *explicit* approaches directly attempt to cover the different query aspects/interpretations/sub-topics. *IA-Select*[Agra09] and *xQuAD*[Sant10] are examples of explicit approaches. In [Varg13] the idea of implicit and explicit approaches is re-visited for *RecSys*. In this case, explicit approaches are those that attempt to cover as many of the user sub-profiles/tastes/preferences as possible.

On the one hand, explicit approaches for *IR* target coverage of query sub-topics to respond to query ambiguity [GilC13]. These approaches assume that: (a) documents that are related to different sub-topics tend to be dissimilar within themselves, and (b) the information that associates a query to all relevant sub-topics is known and complete (e.g., in [Agra09] it is assumed that there is a known distribution $p(c|q)$ that relates categories to queries). In this manner, in *IR*, covering the different sub-topics of the query will result in a list of both diverse and relevant items. Similarly, in *RecSys* it is assumed that the user profile contains information on the diverse tastes/preferences of the user, and that by covering these preferences the final recommendation list will be diverse as well. However, in *RecSys* it is the general case that the information on user preferences is incomplete and available for only a small subset of item categories. Thus, explicit approaches in *RecSys* serve to exploit known information about the user, but on their own they are not enough. In first place, if the user profile is incomplete and contains information only of a subset of homogenous items, then the final recommendation list will not be diverse enough. In second place, explicit approaches penalize novel products and exploration (view *section 2.3.2.2* and *2.3.3*), which imply selecting products that are not related to the user profile.

On the other hand, implicit approaches prevent redundancy within the results by selecting dissimilar items. This poses an advantage over explicit approaches since implicit approaches do not depend on the heterogeneity of the sub-topics/user preferences to produce a diversified list. An additional advantage is that explicit approaches have to develop a strategy to update the marginal utility of query sub-topics to represent how much each sub-topic has already been covered by previously selected documents [GilC13]. However, one major disadvantage of an implicit approach is precisely its independence of the query or user profile. Explicit approaches not only model query sub-topics but also the associated importance/relevance of each sub-topic to the query. This information is useful in order to bias the selection towards more important sub-topics [Zhen12]. In comparison, implicit approaches could be influenced towards items that are very dissimilar from those previously selected but that do not represent an important sub-topic. This is a major disadvantage in *IR*. In *RecSys*, this can actually help the inclusion of novel products promoting exploration. Thus, implicit approaches allow for both exploration and exploitation. However, there is no control over the balance between the exploration vs. exploitation trade-off, potentially resulting in too much exploitation or too much exploration.

In summary, in *RecSys*, explicit approaches exploit the user profile but penalize novel items. In addition, the level of diversity of the final result list would depend on the diversity of the user preferences within the user profile. Alternatively, in implicit approaches redundancy is explicitly prevented. Nonetheless, there is no control over exploitation or exploration, therefore novel products might be added but there is no certainty.

(c) *Control of diversity vs. relevance trade-off*

As has been thoroughly discussed throughout *sections 2.2* and *2.3*, diversification techniques for both *IR* and *RecSys* must offer the correct balance between the diversity and relevance of

generated results. However, because this trade-off could depend on external factors, particularly in the case of *RecSys* as explained in *section 2.3.3*, it is important to offer control parameters to adjust the diversification technique according to the use case requirements of diversity and relevance.

(d) *Encourages Discovery*

Discovery is an important characteristic of *RecSys* as has been discussed in *section 2.3.2.2*. It is important to highlight the capacity of the diversification technique to encourage discovery by not penalizing the inclusion of novel items in the result list.

(e) *Control of exploitation vs. exploration trade-off*

The diversification problem in *section 2.3.3* defines not only the importance of the diversity vs. relevance trade-off but also the importance of the exploitation vs. exploration trade-off. Control over this trade-off would allow the diversification technique to adjust to the different situations that influence positive outcomes, *e.g.*, the user's openness to experience, the heterogeneity of the user profile, among others.

(f) *Considers overlapping item categories*

In most of the discussed approaches, items/documents are classified into a number of categories which are structured according to a domain taxonomy. In this taxonomy defined categories are disjoint and independent from each other. In *IR*, a query is associated to a group of sub-topics/interpretations/categories; and in *RecSys*, user preferences are related to the categories that items within the user profile belong to. In this fashion, preferences are also viewed as disjoint. However, categories might not be disjoint in all domains and thus user preferences could also be overlapping.

For example, in the movie domain, where movies are classified in to genres, the type of movie could be determined by the combination of its genres: a movie belonging to the genres “*Drama Action*” might be very different than a movie belonging to genres “*Drama Thriller*” even if they are both categorized as “*Drama*”. In comparison, a movie categorized as “*Western Action*” might be similar to a movie categorized as “*Western Comedy*”. In an analysis of movie genres, Vargas *et al.* [Varg14] find that certain genres are more general than others: the genre “*Drama*” represents a broad diverse group of movies while in contrast “*Western*” defines a more specific set of movies associated to stories of American Wild West. In their work, Vargas *et al.* [Varg14] concludes genres are not disjoint and they interact with each other. In this manner, it can be deduced that user preferences could also be overlapping: a user that likes “*Western Comedy*” might not like all movies categorized as just “*Comedy*” or just “*Western*”, but mostly the combination of both genres.

Vargas *et al.* [Varg14] point out that current explicit diversification approaches tend to miss the interdependent associations between genres. Moreover, because these approaches attempt to cover as many categories that are in the user profile as possible, items that have many associated categories are favored by the selection process. Nevertheless, it is questionable that an item that covers multiple genres adds equal diversity as the same multiple genres being separately covered by multiple items; for example, diversity ({ (“*Drama, Action, Western*”) }) would be acknowledged as providing equal value as diversity ({ (“*Drama*”); (“*Action*”); (“*Western*”) } [Varg14].

With the defined criteria, the related works are compared in *Table 4*.

	Information Retrieval				Recommendation Systems				
	[Carb98]	[Agra09]	[Sant10]	[Zhen12]	[Zieg05]	[Smyt01]	[Adom09]	[Varg12]	[Shi12]
Greedy Optimization	+	+	+	+	+	+	-	+	+
Explicit Approach	-	+	+	+	-	-	-	+	+
Implicit Approach	+	-	-	-	+	+	-	-	-
Control of diversity vs. relevance trade-off	+	-	+	+	+	+	?	?	+
Encourages Discovery	?	-	-	-	?	?	?	-	-
Control of exploitation vs. exploration trade-off	-	-	-	-	-	-	-	-	-
Considers overlapping item categories	?	-	-	-	?	?	?	-	+

Table 4. Comparison of Diversification Techniques

From Table 4 it can be concluded that:

- Most approaches are based on greedy optimization, which performs very well when the underlying objective function is submodular.
- Approaches are explicit or implicit but not a combination of both.
- None of the approaches explicitly considers not penalizing novel products. However, by chance, some of the approaches, such as implicit ones, could add novel items in the final diversified list.
- None of the approaches consider the trade-off between exploitation and exploration. This can be explained because most approaches for *RecSys* tend to be an adaptation from approaches in *IR* and therefore share the same characteristics. In *IR*, encouraging discovery and exploration are not important factors.
- Most approaches, especially explicit approaches, assume that categories that describe items are disjoint.

In this section we have presented and analyzed diversification techniques for both *IR* and *RecSys*. In the following section diversification evaluation techniques will be analyzed.

3.3 Evaluation Metrics for Diversity

In this section, we present evaluation metrics proposed for both *IR* and *RecSys*, which measure the aspect of diversity within the context of additional factors such as rank, relevance and novelty/redundancy. It is important to highlight that for work in *IR*, the term novelty is mostly related to penalizing redundancy, and thus a document is novel if it contains information that previously selected documents do not have (*i.e.*, non-redundant information). Lastly, we will present a comparative analysis on reviewed metrics in section 3.4.

3.3.1 Evaluation metrics for sub-topic retrieval [Zhai03]

Zhai *et al.* [Zhai03] propose a framework to evaluate sub-topic retrieval, where the main goal is to find documents that cover as many sub-topics as possible of a broader topic. One of the proposed metrics is *sub-topic recall (S-recall) at rank K* defined in Equation 27. For this

equation, it is established there is a broad topic T with n_A associated sub-topics and a ranking of m documents (i.e., d_1, d_2, \dots, d_m). Also, $subtopics(d_i)$ is the set of sub-topics where d_i is relevant. In this manner, S_recall is the percentage of distinct sub-topics covered by at least one document in the first K of the ranking.

$$S_recall \text{ at } K = \frac{|\cup_{i=1}^K subtopics(d_i)|}{n_A} \quad (27)$$

Equation 27. Sub-topic recall at rank K [Zhai03]

The authors clarify that *sub-topic recall* does not offer a meaningful measure across different topics (considering different topics have a different number of associated sub-topics). To solve this, the authors propose two complementary metrics, *S-precision* and *WS-precision* (which penalizes redundancy), that quantify the “intrinsic difficulty” of ranking documents associated to a specific topic. However, solving for both *s-precision* and *WS-precision* require the computation of an *NP-hard* problem to solve for needed optimal values. Furthermore, it can be observed that *S-recall* does not consider factors such as relevance and the position of a document within the result list.

3.3.2 Intent-Aware Evaluation Measures [Agra09][Varg12]

Agrawal et al. [Agra09] explain that “classical IR metrics focus solely on the relevance of documents”. However, metrics should also be intent-aware to account for the fact that certain intents respond better to a given query. The authors propose the following example: We have $P(q|c_2) \gg P(q|c_1)$ and two documents such as $P(d_1|c_1) = 1$ and $P(d_2|c_2) = 1$. If d_1 is rated *Excellent* and d_2 is rated *Good*, then a common IR metric would offer a higher score to the ranking d_1, d_2 even though the query q is most related to the intent c_2 and therefore better responded by a ranking d_2, d_1 .

Two examples of how Agrawal et al. [Agra09] adapt IR metrics to be intent-aware can be viewed in *Table 5*. In the IR intent-aware metric, if a document does not match the intent c it will be judged as “not relevant”, if it does match the intent the document’s relevance score is used. Agrawal *et al.* highlight that the intent-aware adaptation “forces a trade-off between adding documents with higher relevance scores and those that cover additional intents” [Agra09].

Vargas [Varg12] shows that the intent-aware IR metrics can be adapted to be used in *RecSys* using the idea of aspect space (view *section 3.1.2.4*). An example can be viewed in *Equation 28*.

$$nDCG_IA = \sum_{a \in \mathcal{A}} P(a|u) \cdot NDCG(u|a) \quad (28)$$

Equation 28. Intent-aware nDCG adapted to RecSys using the concept of aspect space [Varg12]

Q: ranked result set, k: rank threshold		
Metric	Classical Definition	Intent-Aware Adaptation
Generic Example of IR Metric (M)		$M_{IA}(Q, k) = \sum_c P(c q) \cdot M(Q, k c)$
Normalized Discounted Cumulative Gain (nDCG)	$DCG(Q, k) = \sum_{j=1}^k \frac{2^{r(j)} - 1}{\log(1 + j)}$ $nDCG(Q, k) = \frac{DCG(Q, k)}{DCG(R, k)}$ <p>Where R is the ideal ranking of the set of documents Q and $r(j)$ is the judgement or rating assigned to the item at rank j.</p>	$nDCG_{IA}(Q, k) = \sum_c P(c q) \cdot nDCG(Q, k c)$ <p>Note that there may be more than one possible ideal ranking of the set of documents Q when the metric is intent-aware.</p>
Mean Reciprocal Rank (MRR)	<p>Reciprocal Rank (RR) is “the inverse of the position of the first relevant document in the ordering.” [Agra09]</p> <p>The MRR “of a query set is the average reciprocal rank of all queries in the query set.” [Agra09]</p>	$MRR_{IA}(Q, k) = \sum_c P(c q) \cdot MRR(Q, k c)$

Table 5. Intent-Aware adaptation of classical IR metrics [Agra09]

3.3.3 α -nDCG [Clar08] and Novelty and rank-biased precision (NRBP) [Clar09]

Clarke *et al.* [Clar08] propose the α -nDCG metric to evaluate both novelty and diversity of retrieved results based on cumulative gain.

In their formulation, a user’s information needs u and the information provided by a document d , are modelled by a set of nuggets (*i.e.*, aspects of a query), where $u \subseteq \mathcal{N}$, $d \subseteq \mathcal{N}$ and $\mathcal{N} = \{n_1, n_2, \dots, n_m\}$ is the space of all nuggets. The probability that the user’s information needs contain the nugget n_i is denoted by $P(n_i \in u)$ and the probability a document contains the nugget n_i is denoted by $P(n_i \in d)$. The probability $P(R = 1|u, d)$ determines how relevant a document is to a user by considering the probability that the document contains at least one nugget that is in the user’s information needs, as in Equation 29. This formulation assumes that information on one nugget is independent on information of others, that is $n_i \in u$ is independent of $n_{i \neq j} \in u$ and $n_i \in d$ is independent of $n_{i \neq j} \in d$.

$$P(R = 1|u, d) = 1 - \prod_{i=1}^{|\mathcal{N}|} (1 - P(n_i \in u) \cdot P(n_i \in d)) \quad (29)$$

Equation 29. Probability a document is relevant to a user [Clar08]

On the one hand, to determine $P(n_i \in d)$ the authors assume that a manual binary judgement is given to each document by a human assessor who determines $J(d, i) = 1$ if the document d contains the nugget n_i and $J(d, i) = 0$ if not. In this manner, $P(n_i \in d)$ is determined as in Equation 30, where α is a constant in the range (0,1] that reflects the possibility of human error.

$$P(n_i \in d) = \begin{cases} \alpha, & \text{if } J(d, i) = 1, \\ 0, & \text{otherwise.} \end{cases} \quad (30)$$

Equation 30. Probability a document contains a nugget [Clar08]

On the other hand, to determine $P(n_i \in u)$, the authors assume that in the absence of knowledge over user preferences all nuggets are independent and equally likely to be relevant. To represent this they establish $P(n_i \in u) = \gamma$ for all i , where γ is a constant. Accordingly, Equation 29 can be updated to Equation 31.

$$P(R = 1|u, d) = 1 - \prod_{i=1}^{|\mathcal{N}|} (1 - \gamma \cdot \alpha \cdot J(d, i)) \quad (31)$$

Equation 31. Probability a document is relevant to a user [Clar08]

However, this information gives light on the relevance of a document independent of others in the list. To account for redundancy, the authors indicate that the relevance of a nugget in position k depends on the nuggets in documents that precede it, and formulate $P(n_i \in u | d_1, \dots, d_{k-1})$ as the probability a user is still interested in a nugget defined by the probability that documents he/she has reviewed do not contain the nugget. In this manner, the probability that a document at rank k is relevant (i.e., $R_k = 1$) is determined by Equation 32, where $r_{i,k-1}$ is the number of documents that up to position $k - 1$ have the nugget n_i .

$$P(R_k = 1|u, d_1, \dots, d_k) = 1 - \prod_{i=1}^{|\mathcal{N}|} (1 - \gamma \cdot \alpha \cdot J(d_k, i) \cdot (1 - \alpha)^{r_{i,k-1}}) \quad (32)$$

$$P(R_k = 1|u, d_1, \dots, d_k) = \gamma \cdot \alpha \cdot \sum_{i=1}^{|\mathcal{N}|} J(d_k, i) \cdot (1 - \alpha)^{r_{i,k-1}}$$

Equation 32. Probability a document at rank k is relevant to a user [Clar08]

Using information on document relevance at a rank k for a user u defined by Equation 32, the authors compute gain vectors to be used with the Normalized Discounted Cumulative Gain ($nDCG$) measure. By dropping the constant $\gamma \cdot \alpha$, the authors define that the gain offered by a document at position k is determined by Equation 33.

$$G[k] = \sum_{i=1}^{|\mathcal{N}|} J(d_k, i) \cdot (1 - \alpha)^{r_{i,k-1}} \quad (33)$$

Equation 33. Gain of a document at rank k for a user [Clar08]

In this manner, discounted cumulative gain is formulated as Equation 34.

$$DCG[k] = \frac{\sum_{j=1}^k G[j]}{\log_2(1 + j)} \quad (34)$$

Equation 34. Discounted Cumulative Gain at rank k [Clar08]

Finally, α - $nDCG$ is determined by Equation 35, where $DCG'[k]$ is the DCG of the ideal ordering which maximizes cumulative gain. Computing the ideal ordering is an NP -complete problem usually approximated with greedy optimization.

$$\alpha\text{-}nDCG [k] = \frac{DCG[k]}{DCG'[k]} \quad (35)$$

Equation 35. α - Normalized Discounted Cumulative Gain at rank k [Clar08]

In a nutshell, α - $nDCG$ rewards documents that present new nuggets and penalizes documents that contain redundant nuggets that have already been covered by documents at higher ranks. The parameter α determines how much redundancy is penalized. In this fashion, when $\alpha = 0$, α - $nDCG$ produces the same results as the standard $nDCG$ metric.

In [Clar09], the authors extend the approach inspired on *rank-biased precision (RBP)*, to propose the measure *novelty- and rank-biased precision (NRBP)*, at rank K , defined in Equation 36. In this equation, the parameter β defines the constant probability a user will continue reading down a list of results, assuming the user is reading documents in order from the top of the list. Similarly, β is viewed as the probability a user will continue reading after finding one relevant document.

$$NRBP[k] = \frac{1 - (1 - \alpha)^\beta}{|\mathcal{N}|} \sum_{k=1}^{\infty} \beta^{k-1} \sum_{i=1}^{|\mathcal{N}|} J(d_k, i) \cdot (1 - \alpha)^{r_{i,k-1}} \quad (36)$$

Equation 36. Novelty- and rank-biased precision (NRBP) [Clar09]

3.3.4 Intra-List Similarity [Zieg05]

Ziegler *et al.* [Zieg05] propose the *Intra-List Similarity (ILS)* metric to measure the diversity of a list as a whole, rather than an aggregate of the evaluation of individual items. The *ILS* metric is meant to complement existing accuracy metrics.

The *ILS* metric is defined as in Equation 37.

$$ILS(\mathbb{P}_{w_i}) = \frac{1}{2} \left(\sum_{b_k \in \mathbb{P}_{w_i}} \sum_{b_e \in \mathbb{P}_{w_i}, b_k \neq b_e} sim(b_k, b_e) \right) \quad (37)$$

Equation 37. Re-ranking Method proposed by [Adom09]

Where \mathbb{P}_{w_i} is a list of items (*e.g.*, recommendation list) and $sim(b_k, b_e)$ is an arbitrary function that measures the similarity between two items. Higher values of *ILS* indicate lower diversity. Ziegler *et al.* point out that an interesting feature of the *ILS* metric is *permutation insensitivity*, that is to say that a rearrangement of the elements in the list would not affect the final *ILS* measurement. However, Vargas [Varg12] points out that “*this can be a considerable limitation as far as users do not necessarily browse down to the end of the list, whereby the order in which items are presented may heavily influence the practical utility of the recommendation*”.

As an additional limitation, it can be observed that the *ILS* measure for lists of different sizes cannot be compared. This is because the *ILS* measure is not normalized but just scaled by a factor.

3.3.5 Vargas and Castells formalization of novelty and diversity metrics [Varg11]

Vargas and Castells [Varg11] observe the absence of a clear common conceptual ground for the evaluation of novelty and diversity. They argue that even though many metrics have been proposed to evaluate diversity and novelty, there is not a clear relationship between these

metrics. Moreover, metrics defined in literature that measure novelty or diversity miss important properties: (i) *Relevance awareness*: do not consider the relevance of items; to measure relevance a separate accuracy metric is used; and (ii) *Rank sensitivity*: ignore the ranking of the item disregarding if items are at the bottom or top of the list. As a solution, they propose a formal metric framework for the definition of novelty and diversity metrics.

The proposed framework roots on three essential concepts, modeled as binary random variables, that characterize relations between users and items: (i) *Discovery*: item is seen by the user, (ii) *Choice*: item is consumed by the user, and (iii) *Relevance*: the user liked the consumed item. These variables are naturally related, e.g., a chosen item must be seen.

The metric scheme is built upon two models: *item novelty model* and *browsing model*. Different assumptions and variants for the models and their parameters unfold on to different novelty and diversity metrics.

The most general metric from the framework is defined in *Equation 38*.

$$m(\mathbb{R}|\theta) = C \sum_{i \in \mathbb{R}} p(\text{choose}|i, u, \mathbb{R}) \text{nov}(i|\theta) \quad (38)$$

Equation 38. General specification of Vargas framework for diversity metrics [Varg11]

Where \mathbb{R} , is the list of recommended items to a user, C is a normalizing constant and θ is a generic contextual variable.

On the one hand, in the formula, $p(\text{choose}|i, u, \mathbb{R})$ represents the *browsing model* and indicates the probability the user u will *choose*/consume the item i when delivered within the list \mathbb{R} . The aim is to make the *browsing model* sensitive to both relevance and rank.

On the other hand, in the formula, $\text{nov}(i|\theta)$ represents the *item novelty model* by measuring the novelty of item i given context θ . Vargas and Castells specify two types of item novelty model:

- (i) *Item popularity*: the authors express that, *item novelty can be defined as the difference between an item and “what has been observed” in some context θ* , where different assumptions for θ would lead to different metrics. In this way, item popularity is a way to measure the probability an item has been seen in a given context.
- (ii) *Distance based*: in this case, item novelty is defined by a *distance function between the item and a context of experience*, where the context of experience θ can be a set of items such as items in the user profile or the set of recommended items \mathbb{R} . In this manner, the authors draw a relationship between item novelty and diversity: “*When a set is diverse, each item is “novel” with respect to the rest of the set*” [Varg11]. As follows, the *item novelty model* incorporates diversity measures.

Different instantiations of both models lead to different metrics. Specifically, the authors highlight that by modifying the *browsing model*, *Equation 38* can be rewritten to be rank-sensitive and relevance-aware as in *Equation 39*.

$$m(\mathbb{R}|\theta) = C \sum_{i_n \in \mathbb{R}} \text{disc}(n) p(\text{rel}|i_n, u) \text{nov}(i_n|\theta) \quad (39)$$

Equation 39. Rank-sensitive and relevance-aware diversity metric [Varg11]

Where $\text{disc}(n)$ is a discount function that considers the position of the item in the recommendation list and $p(\text{rel}|i_n, u)$ represents the relevance of the item.

The authors show that their framework unifies and generalizes several state-of-the-art metrics. Also, they propose mechanisms to enhance existing metrics with configurable properties that consider relevance awareness and rank sensitivity. In addition, through offline experimental validation, Vargas and Castells conclude that a unified metric that considers both novelty/diversity and relevance at the same time, is better than a combination of two separate metrics (*i.e.*, a combination of an accuracy metric and a novelty/diversity metric).

3.4 Discussion on Evaluation Metrics for Diversity

In this section, we have presented metrics that evaluate diversity in the fields of *IR* and *RecSys*. On the one hand, for the field of *IR*, intent-aware metrics and α -*nDCG* have been official metrics for the diversity task in the TREC Web Track³, and therefore have been used to evaluate a great part of the current work on *IR* diversification techniques. On the other hand, for the field of *RecSys*, intra-list similarity (*ILS*) and the evaluation framework proposed by Vargas and Castells (view *section 3.3.5*) are representative of the metrics used to evaluate diversification techniques in *RecSys*.

To compare the reviewed metrics, we establish comparison criteria in *Table 6*. First and foremost, criteria must allow to analyze metrics in accordance to the trade-offs we have defined compose the *RecSys* diversification problem: diversity *vs.* relevance and exploitation *vs.* exploration (view *section 2.3.3*). It is important to highlight that we are not looking for a metric that meets all the defined characteristics, given the difficulty of defining an integrated metric. However, it would be ideal to identify which metrics can evaluate different aspects of the diversification techniques. In addition, metrics should consider the overlapping characteristic of item features and user preferences that has been previously explained in *section 3.2*.

Criteria	Definition
<i>Relevance-awareness</i>	Is the metric relevance-aware?
<i>Rank-awareness</i>	Does the metric take into consideration the order of items?
<i>Penalizes Redundancy</i>	Does the metric explicitly penalize redundancy within the results?
<i>Measures Exploitation</i>	Can the metric evaluate if the different user-preferences are covered?
<i>Measures Exploration</i>	Can the metric evaluate novelty/serendipity of results?
<i>Considers overlapping item categories</i>	Does the metric consider that item characteristics could be interdependent?
<i>Comparable among result lists of different sizes</i>	Can the results of the metric be compared if the size of the result list changes?
<i>Measures diversity</i>	Can the metric measure at least one of the properties of diversity (<i>i.e.</i> , disparity, balance and variety)?

Table 6. Comparison Criteria for Metrics to Evaluate Diversity

³ TREC Web Track Guidelines: <http://plg.uwaterloo.ca/~trecweb/>. Last viewed: November 23, 2014.

With the defined criteria, the reviewed metrics are compared in *Table 7*.

	S-recall [Zhai03]	Intent-Aware Metrics [Agra09][Varg12]	α -nDCG and NRBP [Clar08][Clar09]	ILS [Zieg05]	Novelty and diversity framework [Varg11]
Relevance-awareness	-	?	+	-	+
Rank-awareness	-	?	+	-	+
Penalizes Redundancy	-	-	+	?	?
Measures Exploitation	-	?	-	-	+
Measures Exploration	-	-	-	-	+
Considers overlapping item categories	-	-	-	+	?
Comparable among result lists of different sizes	-	+	?	-	+
Measures diversity	+	-	?	+	?

Table 7. Comparison of Metrics to Evaluate Diversity

From analysing metrics we can conclude that:

- Intent-aware metrics can be relevance-aware or rank-aware depending on the metric being adapted to be intent aware (*e.g.*, *IA-nDCG* would be both relevance and rank aware).
- Most metrics do not explicitly penalize redundancy within the retrieved results.
- Intent-aware metrics do consider aspects of the user profile, where aspects are defined as proposed by [Varg12]. In this manner, these metrics exploit known information about the user. However, the metrics give priority to the most important aspects or preferences of the user profile, and therefore covering aspects that are not as important represent less gain. In this manner, these metrics encourage diversity only if the user's most important preferences are diverse.
- The α -nDCG metric penalizes redundancy, in addition to being relevance and rank aware. However, to derive this metric a number of strong assumptions are made. In first place, the metric assumes that nuggets are independent of each other, and therefore the metric cannot consider possible overlapping of item categories or user preferences. Moreover, the metric assumes there is no information on user preferences, and therefore does not exploit known information within the user profile. These assumptions might be reasonable in the field of *IR* but not for the field of *RecSys*.
- The only work that considers the notion of novelty as defined in the field of *RecSys* (*i.e.*, as defined in *section 2.3.2.2*) is the Vargas *et al.* [Varg11] framework.
- The only metric that truly considers that item categories can be overlapping is the *ILS* metric as it uses similarity measures.
- Most of the metrics do not explicitly consider diversity.

All in all, we observe that a good metric to measure the aspect of diversity is the *IntraList Similarity (ILS)* metric. Also, we can extract from the Vargas *et al.* [Varg11] framework the foundations for metrics that consider novelty, rank and relevance.

In this section, we have presented an analyzed metrics to evaluate diversity for both *IR* and *RecSys*. In the following section, a brief summary of the literature review chapter is offered.

3.5 Summary

In this chapter we have analyzed related works on both diversification techniques and diversity-aware metrics for the fields of *IR* and *RecSys*. We find that current works for *RecSys* are pure adaptations of solutions from *IR*, and as such, miss important properties essential to *RecSys*. We highlight that even though *IR* works can serve as a foundation for advances in *RecSys*, the ultimate goal of *RecSys* is very different from that of *IR*. Consequently, it is not enough to just accommodate *IR* ideas towards *RecSys*, they must also be augmented to consider characteristics fundamental to *RecSys* such as novelty and discovery. Towards this goal, in the following chapter we will introduce the Exploitation-Exploration Diversification Technique. This technique, not only considers the trade-off between relevance and diversity, but also the trade-off between exploitation of the user profile and exploration of novel products. As a result, the approach can be tuned towards either more explorative or more exploitative recommendations depending on the characteristics of the user and the product domain.

Chapter IV

EXPLOITATION-EXPLORATION DIVERSIFICATION TECHNIQUE

In this chapter we will introduce the *Exploitation-Exploration Diversification* technique named *XPLODIV*, which is a post-filtering approach that follows the structure of *Figure 5*, where a subset of diversified items is selected from a list of candidate items generated from a traditional *RecSys*. We formulate our approach as a greedy optimization problem that aims to retrieve an ordered subset of items \mathbb{R} , by iteratively adding to \mathbb{R} one item i^* from the set of candidate items \mathbb{C} , where i^* maximizes the function *XPLODIV*—defined in *Equation 40*— at a given iteration step, as shown in the algorithm in *Figure 9*. The technique receives as input the user profile \mathbb{U} of the target user (*i.e.*, the set of items $u \in \mathbb{U}$ that the target user has rated), a set of candidate items \mathbb{C} —specifically generated for the particular target user— and the desired size for \mathbb{R} defined as k . As output, *XPLODIV* produces an ordered set of diversified items \mathbb{R} , where $\mathbb{R} \subseteq \mathbb{C}$ and $|\mathbb{R}| = k$, to be presented as the final recommendation list.

Input: output size k , set of user profile items \mathbb{U} , set of candidate items \mathbb{C}
Output: set of diversified items \mathbb{R}

- 1: $\mathbb{R} \leftarrow \emptyset$
- 2: **while** $|\mathbb{R}| < k \wedge \mathbb{C} \neq \emptyset$ **do**
- 3: $i^* \leftarrow \arg \max_{i \in \mathbb{C} \setminus \mathbb{R}} \text{XPLODIV}(i, \mathbb{U}, \mathbb{R})$
- 4: $\mathbb{C} \leftarrow \mathbb{C} \setminus \{i^*\}$
- 5: $\mathbb{R} \leftarrow \mathbb{R} \cup \{i^*\}$
- 6: **end while**
- 7: **return** \mathbb{R}

Figure 9. XPLODIV Greedy Optimization Algorithm

$$\text{XPLODIV}(i, \mathbb{U}, \mathbb{R}) = \alpha \cdot \text{rel}(i) + (1 - \alpha) \cdot \text{div}(i, \mathbb{R}) \cdot (\beta \cdot \text{exploit}(i, \mathbb{U}) + (1 - \beta) \cdot \text{xplore}(i, \mathbb{U})) \quad (40)$$

Equation 40. Exploitation-Exploration diversification XPLODIV

As was mentioned in *section 2.3.3*, the goal of *RecSys* diversification is to balance the trade-off between *relevance* and *diversity*, considering the trade-off between *exploitation* of the user profile and *exploration* of novel products. To achieve this, *XPLODIV* has four core dimensions:

- *Relevance* — $\text{rel}(i)$
- *Diversity* — $\text{div}(i, \mathbb{R})$
- *Exploitation* — $\text{exploit}(i, \mathbb{U})$
- *Exploration* — $\text{xplore}(i, \mathbb{U})$

Moreover, the approach has two control parameters:

- (i) The parameter α controls the trade-off between relevance and diversity.
- (ii) The parameter β controls the trade-off between exploitation and exploration.

Lastly, an important aspect to highlight is that the diversity of selected items in *XPLODIV* is directly linked to the *exploitation vs. exploration* trade-off. As a result, the approach can be set towards more diverse exploitative items or more diverse explorative items.

A concise description of the four dimensions that compose *XPLODIV* is offered in *Table 8*. Each of these must be normalized to return a value in the range $[0,1]$, where one indicates the highest desirable value. The remainder of this chapter will discuss in detail the dimensions that compose *XPLODIV* and possible specifications for each dimension. Lastly, a concise summary is presented.

Dimension	Function	Definition
Relevance	$rel(i)$	Returns the normalized predicted rating of item i obtained as a result of the traditional <i>RecSys</i> .
Diversity	$div(i, \mathbb{R})$	Returns a measure of diversity between item i and the set of previously selected items \mathbb{R} .
Exploitation	$xploit(i, \mathbb{U})$	Returns the exploitation value of item i with respect to the user profile \mathbb{U} .
Exploration	$xplore(i, \mathbb{U})$	Returns the exploration value of item i with respect to the user profile \mathbb{U} .

Table 8. XPLODIV Dimensions

Relevance Dimension

The relevance dimension gives priority to items that have high predicted rating. In this fashion, the relevance value of an item i is given by *Equation 41*. In this equation, *MaxRating* is the maximum possible rating that a user can give to an item and *predicted_rating(i)* is the predicted rating for item i obtained from a traditional *RecSys*.

$$rel(i) = \frac{predicted_rating(i)}{MaxRating} \quad (41)$$

Equation 41. Relevance Dimension

Diversity Dimension

The diversity dimension $div(i, \mathbb{R})$ measures how diverse an item i is in relation to a set of items \mathbb{R} . In order to measure the diversity of one item to a set we can use *Equation 3a*, which defines $div(i, \mathbb{R})$ in terms of the amount of diversity that would be gained for the set \mathbb{R} if a new element i is added. In this manner, any measure of set diversity can be used (e.g., Gini-Simpson diversity index, Stirling diversity, among others). For example, in *Equation 42* we can view how to use *Equation 3a* with Stirling's measure of diversity defined in *Equation 4*.

$$\begin{aligned} div(i, \mathbb{R}) &= diversity(\mathbb{R} \cup \{i\}) - diversity(\mathbb{R}) \\ diversity(\mathbb{R}) &= \Delta_{diversity}(\mathbb{R}) \end{aligned} \quad (42)$$

Equation 42. Example of Diversity that an element i would add to a set \mathbb{R}

Alternatively, according to Weitzman [Weit92], the diversity item i would add to the set of items \mathbb{R} is determined by the distance of the item to the set. Weitzman [Weit92] specifies the minimum distance to be one measure, as in *Equation 3b*. Another method to measure the item-

to-set distance can be defined as the average pairwise distance of the item i to each of those items in the set as in *Equation 43*. The distance measure $\text{distance}(i, r)$ can derive from a similarity function as in *Equation 6*.

$$\text{div}(i, \mathbb{R}) = \frac{1}{|\mathbb{R}|} \sum_{r \in \mathbb{R}} \text{distance}(i, r) \quad (43)$$

Equation 43. Average pairwise distance of an element i to a set \mathbb{R}

Exploitation Dimension

The exploitation dimension wishes to reinforce those items that exploit known user preference information. These items are those that are representative of the user's tastes found within the user profile. Items that are close to previously identified user preferences could turn out to be promising recommendations, following the content-based *RecSys* heuristic that assumes users will continue to have the same preferences they have had in the past.

To achieve this, the exploitation dimension aims to determine how representative item i is of the user's preferences found in the user profile \mathbb{U} of the target user. In related works discussed in *section 3.1*, measures of the exploitation dimension can be found in explicit approaches, where a model of the importance of query sub-topics is used to find those documents that cover the most important sub-topics. However, these approaches are not suitable for *RecSys* as they do not consider that user preferences can be overlapping. We propose two approaches that use similarity measures between the items in order to capture the representativeness of the underlying user preferences. The use of similarity metrics instead of partitioning the user profile in to rigid preferences better responds to the implicit overlapping user preferences without the need of a limiting categorization.

One way to determine the exploitation value of an item is to calculate the probability that similar items within the user profile have a high rating, as in *Equation 44*. This equation calculates the sum of the target user's ratings weighted by the similarity each item in the user profile has to the item i , normalized by the overall sum of user ratings. In *Equation 44* the function $\text{rating}(u, \mathbb{U})$ returns the rating the user assigned to the item u .

$$\text{xexploit}(i, \mathbb{U}) = \frac{\sum_{u \in \mathbb{U}} \text{sim}(i, u) \cdot \text{rating}(u, \mathbb{U})}{\sum_{u \in \mathbb{U}} \text{rating}(u, \mathbb{U})} \quad (44)$$

Equation 44. Exploitation dimension as probability of high rating of similar items

An extension to this approach is to determine the probability the nearest neighbors of the item within the user profile have a high rating, as in *Equation 45*. Nearest neighbors can be chosen given a size of k nearest neighbors or given a similarity threshold.

$$\text{xexploit}(i, \mathbb{U}) = \frac{\sum_{n \in \text{nearestNeighbors}(i, \mathbb{U})} \text{rating}(n, \mathbb{U})}{\sum_{u \in \mathbb{U}} \text{rating}(u, \mathbb{U})} \quad (45)$$

Equation 45. Exploitation dimension as probability of high rating of nearest neighbors

The exploitation dimension can be extended to consider additional factors. For example, in certain domains where recommendations depend on context, the exploitation dimension could just exploit those portions of the user profile that are the most relevant to the given context. In order to develop these measures for the exploitation dimension, designers must answer questions such as: which portions of the user profile are relevant given a determined

context/user intention/time-frame?. However, answers to these questions are out of the scope of this project and left for future work.

Exploration Dimension

The exploration dimension wishes to reinforce those items that allow the user to discover and explore the unknown. In other words, the exploration dimension would give priority to novel/serendipitous items that are outside of the user's past tastes. Given that the user profile can be ambiguous and incomplete, it is not smart to always exploit known information and possibly stay stuck in a sub-optimal item space. By offering user's novel products, the *RecSys* is also attempting to retrieve information on unknown user preferences, preventing overspecialization.

In *section 2.3.2.2*, we concluded novelty could be measured as how diverse an item is from the user's past experiences. In this manner, we can use one of the specified measures for the diversity dimension; but instead of measuring the diversity of item i to the list of selected items \mathbb{R} , we measure the diversity of item i in relation to the target user's past experiences. The most clear indication of the user's past experiences is encapsulated within the user profile \mathbb{U} . Hence, exploration could be measured as in *Equation 46*.

$$xplore(i, \mathbb{U}) = div(i, \mathbb{U}) \quad (46)$$

Equation 46. Exploration as diversity of item i to the user profile \mathbb{U}

Another way to determine the user's past experiences could be by considering the experiences of similar users. In other words, we could assume the target user has similar experiences to other users with similar preferences. For example, if an item is well known or popular among users that are similar to the target user, it is probable that the target user already knows about this product even though he/she has not rated the item yet. In this case, we would like to measure the novelty of the item with respect to the neighborhood of similar users, in addition to the user profile, as in *Equation 47*. In this equation, the profiles of the k nearest neighbors of the user are aggregated in to a set \mathbb{N} . Alternatively, a similarity threshold can be used to obtain the nearest neighbors. Next, the exploration dimension is determined as the diversity of item i to the set of items formed by the union of the user profile and the profiles of the nearest neighbors.

$$\begin{aligned} \mathbb{N} &= \{\mathbb{N}_1 \cup \mathbb{N}_2 \dots \cup \mathbb{N}_k\} \\ xplore(i, \{\mathbb{N} \cup \mathbb{U}\}) &= div(i, \{\mathbb{N} \cup \mathbb{U}\}) \end{aligned} \quad (47)$$

Equation 47. Exploration as diversity of item i to the nearest neighbors

Summary

In this section, we have defined the Exploitation-Exploration diversification technique *XPLODIV*. As an advantage over current works, *XPLODIV* does not only consider the trade-off between relevance and diversity but also the trade-off between exploration and exploitation. In this manner, *XPLODIV* can be adjusted towards more or less exploitative or explorative recommendations according to the requirements of the *RecSys*. In the following chapter experimental validation of *XPLODIV* is offered.

Chapter V

EXPERIMENTAL VALIDATION

The aim of this chapter is to show, through experimental validation, that the Exploitation-Exploration diversification technique named *XPLODIV* satisfies the following hypothesis:

- (i) *Hypothesis I*: the *XPLODIV* approach can be tuned towards different configurations of relevance, diversity, exploitation and exploration.
- (ii) *Hypothesis II*: the *XPLODIV* approach produces results comparable to baseline techniques in terms of relevance and diversity.

With the proposed experiments we want to answer the following questions:

- Can the trade-offs (*i.e.*, relevance *vs.* diversity and exploitation *vs.* exploration) be observed?
- Can the parameters α and β control the characteristics of generated results in terms of relevance, diversity, exploitation and exploration?
- In which scenarios does *XPLODIV* outperform baseline techniques?
- In which scenarios does *XPLODIV* perform worse than baselines techniques?
- Do different configurations of *XPLODIV* affect outcomes?
- What additional trade-offs can be observed? Is there a relation between the following properties: exploitation and diversity, exploitation and relevance, exploration and diversity, exploration and relevance?

We have carried out two types of offline experiments over the MovieLens 100k [Grou14] dataset: (i) *qualitative*, which are presented in *section 5.2*; and (ii) *quantitative*, which are presented in *section 5.3*. Qualitative tests aim to prove that the *XPLODIV* approach can be tuned towards different configurations of exploitation and exploration, therefore offering partial proof for *Hypothesis I*. Quantitative tests are more detailed and aim to prove that *Hypothesis I* and *II* are true.

The overall evaluation environment is described in *section 5.1*, where the tested techniques and their configurations are presented. In addition, for both types of tests, we analyzed results for user profiles with different levels of heterogeneity, where heterogeneity depends on the number of unique item genres that can be found in the user profile (*e.g.*, users who have rated items from few movie genres have a homogeneous user profile). We believe that user profile heterogeneity has an important impact on diversification results, which is why this aspect is relevant for all tests. The calculation of user profile heterogeneity is further explained in *section 5.1*.

It is important to clarify that for the context of the current project we want to observe if *XPLODIV* can perform as well as baseline techniques in terms of diversity and relevance, and that in addition, results can be tuned using the control parameters. For that reason, tests with users are out of the scope of our project, as our goal is not to evaluate the impact of producing results with different characteristics (such as diversity, exploitation and exploration) on user satisfaction. In brief, we want to show that our technique performs as expected.

Finally, in *section 5.4* a summary of findings is presented.

5.1 Evaluation Environment

We carry out tests over the MovieLens 100k [Grou14] dataset, which has 100,000 ratings (rating scores are in the range [1,5]) from 943 users on 1682 movies, where each user has rated at least 20 movies. Data is provided by the MovieLens website which is maintained by the GroupLens Research Project at the University of Minnesota.

Our approach is compared to the following baselines and state-of-the-art techniques:

- (a) *No Diversity*: returns the top k of candidate items.
- (b) *Random Diversity*: returns a random selection of k items from candidate items.
- (c) *MMR* with $\alpha = 0.5$ (refer to *section 3.1.1.1*): returns k items selected with the technique *MMR*, which is a representative of current implicit *RecSys* diversification approaches.

Explicit diversification approaches are purposely omitted as by definition they are biased towards only exploitative items ignoring novel/explorative items.

The set of candidate items, sorted in descending order according to the predicted rating, is generated from a traditional user-user collaborative filtering *RecSys* with user neighborhood size of 50^4 . We choose a user-based collaborative filtering algorithm given that these are less prone to overspecialization problems compared to content-based algorithms. To measure item similarity, we used the Jaccard coefficient, defined as $Jaccard(A, B) = |A \cap B| / |A \cup B|$, between the set of movie genres associated to the items being compared. The same set of candidate items, generated for a particular target user, serves as input for all the diversification techniques under evaluation for that user. We fixed the size of the set of candidate items to 100. Through empirical observation, we found that a larger set of candidate items would not significantly impact results in the case of the MovieLens 100k dataset.

Diversification techniques were implemented in Java, using the Apache Mahout machine learning library [Apac14]. Apache Mahout provides a *RecSys* framework with ready to use components that facilitate tasks such neighborhood selection, similarity calculations and data model interaction (*i.e.*, interaction with the User-Item Matrix from *Figure 4*). Also, the framework provides configurable user and item based collaborative filtering algorithms; specifically, we rely on the Mahout user-user collaborative filtering recommender to generate candidate items. Other tools that were used are: Apache Maven for project management and JUnit as a testing framework.

In *Figure 10*, a class diagram of the implementation of diversification techniques is presented. In our design, selected diversification techniques can be configured to use different implementations of the exploitation, exploration and diversity dimensions. The implementation of these dimensions is corresponding with those explained in *chapter IV* and are presented in *Table 9*.

Specifically, *MMR* can be configured with an instantiation of the *DiversityDimension* component. By default, *MMR* uses maximum similarity (view *Equation 9*), therefore in our implementation *MMR* uses by default the *MinimumDissimilarity* component as diversity measurement. Lastly, our approach is implemented in the class *ExploreExploitDiversification* which can be configured with different combinations of the types of dimensions, represented by the abstract classes *DiversityDimension*, *ExplorationDimension* and *ExploitationDimension*.

⁴ Further details on the user-based collaborative filtering algorithm provided by Mahout can be found at [Owen12].

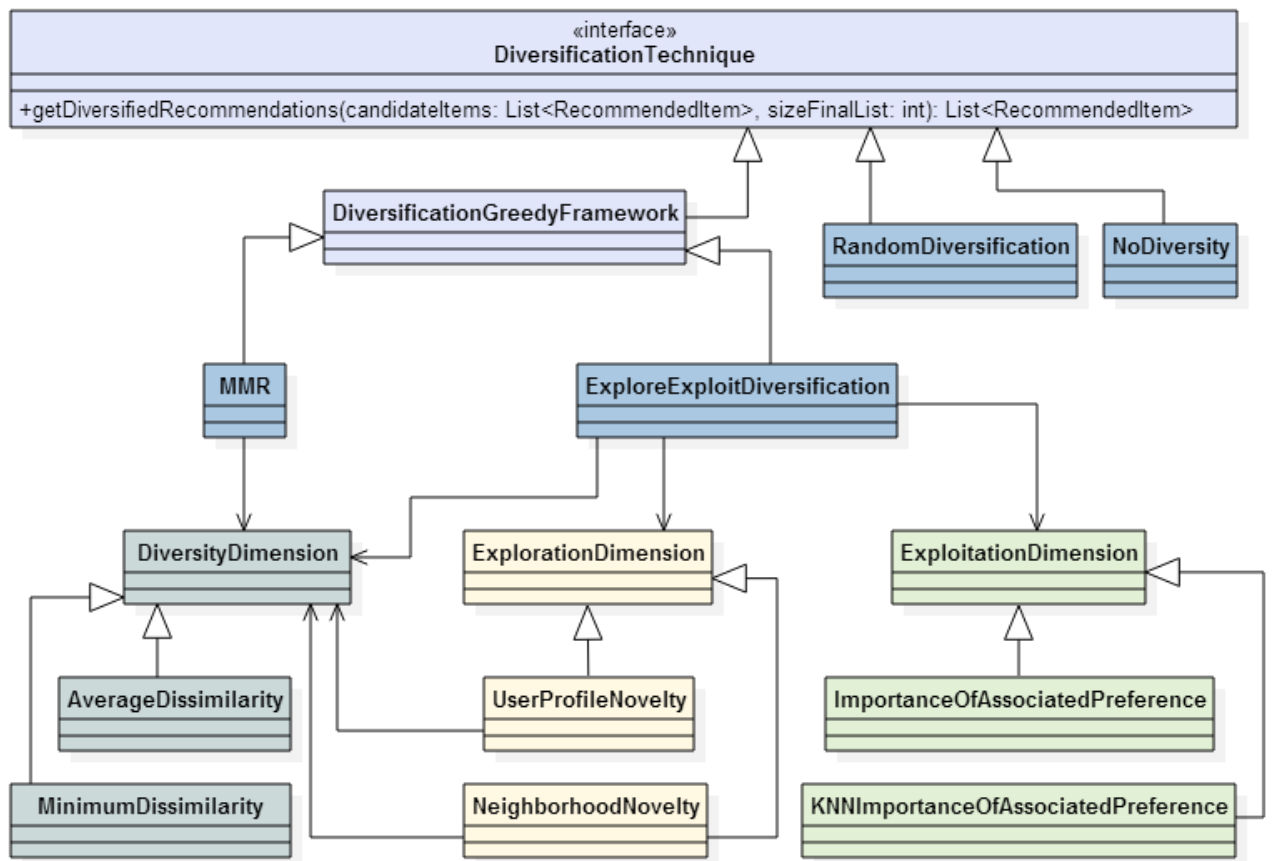


Figure 10. Diversification Techniques Class Diagram

Class Name	Dimension	Definition
<i>AverageDissimilarity</i>	Diversity	View Equation 43
<i>MinimumDissimilarity</i>	Diversity	View Equation 3b
<i>UserProfileNovelty</i>	Exploration	View Equation 46
<i>NeighborhoodNovelty</i>	Exploration	View Equation 47
<i>ImportanceOfAssociatedPreference</i>	Exploitation	View Equation 44
<i>KNNImportanceOfAssociatedPreference</i>	Exploitation	View Equation 45

Table 9. Definition of Implemented Dimensions

In the following sections, data interpretations for both qualitative and quantitative tests are presented. Tests were carried out considering the level of heterogeneity of user profiles. As a consequence, we were able to observe the influence that the diversity of the user's preferences has on diversification results. For the MovieLens dataset, we propose as a measurement for the level of heterogeneity of a user profile, the percentage of unique genres that can be found associated to items in the user profile; as defined in Equation 48.

$$UserProfileHeterogeneity = \frac{UniqueGenresInUserProfile}{TotalNumberOfGenresInMovieLens} * 100 \quad (48)$$

Equation 48. User Profile Heterogeneity for MovieLens dataset

5.2 Qualitative Offline Experiment

With qualitative tests, we want to observe the tunability of *XPLODIV*, and analyze the effects of using different instantiations of the dimension components. In order to do so, we define two general test scenarios: (i) *pure exploitation*: $\alpha = 0.0$, $\beta = 1.0$; and (ii) *pure exploration*: $\alpha = 0.0$, $\beta = 0.0$. In *pure exploitation*, results should have items similar to those in the user profile, and in *pure exploration*, results should have items as far away from the user profile as possible. We set $\alpha = 0.0$ for both scenarios, as observations of the relevance aspect will be carried out in the quantitative tests. Concisely, with qualitative tests we would like to observe that results can be tuned to be exclusively explorative or exclusively exploitative, thus presenting evidence to partially prove *Hypothesis I*.

For both scenarios, we analyze results for three users which differ in their associated levels of heterogeneity —grounded on the number of unique genres found in the user profile—, as follows:

- (a) *Low heterogeneity*: represented by *user 914* with heterogeneity level of 31.5%.
- (b) *Medium heterogeneity*: represented by *user 53* with heterogeneity level of 53,5%
- (c) *High heterogeneity*: represented by *user 96* with heterogeneity level of 94.7%.

We choose these users as they have small user profiles (composed of less than 60 items), making it plausible to visually compare the items within the set of results to the items within the user profile.

For all qualitative experiments we used $k=10$, *i.e.*, the generated output would be a set of ten items. This size is common in *RecSys* applications, and also, it is small enough to visually analyze results.

In the following sections we discuss results from both test scenarios.

5.2.1 Discussion on Pure Exploitation Results

In the scenario of *pure exploitation*, *XPLODIV* parameters are set to $\alpha = 0.0$ and $\beta = 1.0$. We run tests for different combinations of the possible instantiations of the exploitation dimension using the diversity dimension, as shown in *Table 10*.

		Exploitation Dimension	
		Importance of Associated Preference (IOAP)	KNN Importance of Associated Preference (KNN IOAP)
Diversity Dimension	Average Dissimilarity (Avg. Diss.)	I IOAP Avg. Diss.	III KNN IOAP Avg. Diss.
	Minimum Dissimilarity (Min. Diss.)	II IOAP Min. Diss.	IV KNN IOAP Min. Diss.

Table 10. Pure exploitation test combinations

Test results are shown in *Table 11*, where the output generated by *XPLODIV* for the different test combinations is shown for each of the users. Detailed observations for all outputs are offered. In general, we observed if genres in the user profile could be found within results and if results presented genres that were not in the user profile. In addition, by observing genre pair

co-appearances found in items within the user profile, we identified the most frequent genre associations as “Relevant genre associations”, to verify if they could be found within generated results. In the Table, the column “*XPLODIV* Output” presents the ranked list of generated results by our approach. For each item in the results, we present the item’s associated genres and within square brackets the item’s identifier. In order to emphasize important observations within results, for each item we: (i) differentiate with green if the genre that is the most frequently rated by the user is found in the genres associated to the item, (ii) highlight with orange the relevant genre pair associations identified and (iii) strikethrough genres that are not found in the user profile and that for exploitation purposes are not desired.

User Profile	Test	<i>XPLODIV</i> Output	Observations														
<p>Low Heterogeneity - User 914 Size of Profile: 23 items</p> <p>Genre appearance frequency:</p> <table border="1"> <thead> <tr> <th>Genre</th> <th>Frequency</th> </tr> </thead> <tbody> <tr> <td>Romance</td> <td>21</td> </tr> <tr> <td>Comedy</td> <td>13</td> </tr> <tr> <td>Drama</td> <td>12</td> </tr> <tr> <td>Musical</td> <td>2</td> </tr> <tr> <td>Action</td> <td>1</td> </tr> <tr> <td>Thriller</td> <td>1</td> </tr> </tbody> </table>	Genre	Frequency	Romance	21	Comedy	13	Drama	12	Musical	2	Action	1	Thriller	1	<p>I IOAP Avg. Diss.</p>	<ol style="list-style-type: none"> [517] Comedy/Drama/Romance [705] Musical/Romance [213] Drama/Romance [487] Comedy/Romance [490] Comedy/Romance/Thriller [875] Drama/Romance [514] Comedy/Romance [131] Drama/Romance [133] Drama/Romance/War [1197] Comedy/Drama 	<ul style="list-style-type: none"> • First item has the three most important genres and both relevant genre associations. • Covers all genres in the user profile except for Action. • Top 9 items all have the most frequent genre, which is Romance • Results cover all the important genre associations. • Only one item is associated to a genre that is not in the profile.
Genre	Frequency																
Romance	21																
Comedy	13																
Drama	12																
Musical	2																
Action	1																
Thriller	1																
<p>Relevant genre associations:</p> <table border="1"> <thead> <tr> <th>Genre Pair</th> <th>Frequency</th> </tr> </thead> <tbody> <tr> <td>Comedy, Romance</td> <td>13</td> </tr> <tr> <td>Drama, Romance</td> <td>10</td> </tr> </tbody> </table>	Genre Pair	Frequency	Comedy, Romance	13	Drama, Romance	10	<p>II IOAP Min. Diss.</p>	<ol style="list-style-type: none"> [517] Comedy/Drama/Romance [705] Musical/Romance [490] Comedy/Romance/Thriller [133] Drama/Romance/War [213] Drama/Romance [487] Comedy/Romance [209] Comedy/Drama/Musical [337] Comedy/Drama/Thriller [165] Drama [185] Horror/Romance/Thriller 	<ul style="list-style-type: none"> • Bottom four items are the only ones that are not within the results of Test I (IOAP Avg. Diss.). First two items are the same as Test I. • Results cover all the important genre associations. • Two items have genres not in the user profile. • Covers all genres in the user profile except for Action. 								
Genre Pair	Frequency																
Comedy, Romance	13																
Drama, Romance	10																
	<p>III KNN IOAP Avg. Diss.</p>	<ol style="list-style-type: none"> [172] Action/ Adventure/ Drama/ Romance/ Sci-Fi/ War [185] Horror/Romance/Thriller [191] Drama/Mystery [337] Comedy/Drama/Thriller [705] Musical/Romance [1194] Crime/Drama [1152] Romance/War [50] Action/Adventure/Romance/Sci-Fi/War [1197] Comedy/Drama [213] Drama/Romance 	<ul style="list-style-type: none"> • Covers all the genres in the user profile. • Seven items are associated to genres that are not in the user profile. • Doesn’t keep as much of the relevant genre associations. • Items appear to have more associated genres. 														

	<p>IV KNN IOAP Min. Diss.</p>	<p>1. [172] Action/ Adventure/ Drama/ Romance/ Sci-Fi/ War 2. [185] Horror/ Romance/ Thriller 3. [191] Drama/ Mystery 4. [95] Animation/ Children's/ Comedy/ Musical 5. [337] Comedy/ Drama/ Thriller 6. [213] Drama/ Romance 7. [127] Action/ Crime/ Drama 8. [1152] Romance/ War 9. [528] Drama/ War 10. [705] Musical/ Romance</p>	<ul style="list-style-type: none"> • Covers all the genres in the user profile. • Only three items are not within the results of Test III (KNN IOAP Avg. Diss.), which are [95],[127],[528]. First three items are the same as Test III. • Doesn't keep as much of the relevant genre associations. • Seven items are associated to genres that are not in the user profile. 																																
<p>Medium Heterogeneity - User 53 Size of Profile: 27 items Genre appearance frequency:</p> <table border="1" data-bbox="240 835 523 1317"> <thead> <tr> <th>Genre</th> <th>Frequency</th> </tr> </thead> <tbody> <tr> <td>Action</td> <td>15</td> </tr> <tr> <td>Adventure</td> <td>9</td> </tr> <tr> <td>Sci-Fi</td> <td>9</td> </tr> <tr> <td>Drama</td> <td>8</td> </tr> <tr> <td>Thriller</td> <td>8</td> </tr> <tr> <td>Comedy</td> <td>5</td> </tr> <tr> <td>Romance</td> <td>5</td> </tr> <tr> <td>Crime</td> <td>4</td> </tr> <tr> <td>War</td> <td>4</td> </tr> <tr> <td>Children's</td> <td>1</td> </tr> </tbody> </table> <p>Relevant genre associations:</p> <table border="1" data-bbox="240 1391 523 1731"> <thead> <tr> <th>Genre Pair</th> <th>Frequency</th> </tr> </thead> <tbody> <tr> <td>Action, Adventure</td> <td>7</td> </tr> <tr> <td>Action, Romance</td> <td>4</td> </tr> <tr> <td>Thriller, Action</td> <td>6</td> </tr> <tr> <td>Sci-Fi, Action</td> <td>6</td> </tr> </tbody> </table>	Genre	Frequency	Action	15	Adventure	9	Sci-Fi	9	Drama	8	Thriller	8	Comedy	5	Romance	5	Crime	4	War	4	Children's	1	Genre Pair	Frequency	Action, Adventure	7	Action, Romance	4	Thriller, Action	6	Sci-Fi, Action	6	<p>I IOAP Avg. Diss.</p> <p>II IOAP Min. Diss.</p> <p>III KNN IOAP Avg. Diss.</p>	<p>1. [298] Action/ Sci-Fi/Thriller 2. [1127] Drama 3. [127] Action/Crime/Drama 4. [689] Action/Thriller 5. [201] Action/ Adventure/ Comedy/ Horror 6. [326] Action/Drama/War 7. [355] Adventure/Sci-Fi/Thriller 8. [300] Action/Thriller 9. [313] Action/Drama/Romance 10. [332] Crime/Drama/Thriller</p> <p>1. [298] Action/Sci-Fi/Thriller 2. [1127] Drama 3. [127] Action/Crime/Drama 4. [201] Action/ Adventure/ Comedy/ Horror 5. [326] Action/Drama/War 6. [4] Action/Comedy/Drama 7. [129] Crime/ Drama/ Romance/ Thriller 8. [313] Action/Drama/Romance 9. [315] Drama/Thriller 10. [355] Adventure/Sci-Fi/Thriller</p> <p>1. [332] Crime/Drama/Thriller 2. [201] Action/ Adventure/ Comedy/ Horror 3. [521] Drama/War 4. [89] Film Noir/Sci-Fi 5. [304] Adventure/Children's 6. [327] Crime/Drama/Mystery 7. [558] Drama/Fantasy/Thriller 8. [510] Action/Drama/Western 9. [408] Animation/Comedy/Thriller 10. [511]Adventure/War</p>	<ul style="list-style-type: none"> • Covers all the relevant genre associations. • Missing only one genre (<i>i.e.</i>, Children's) to cover all genres in the user profile. • First item has two of the relevant genre associations. • Most items contain the most frequent genre, which is Action. • Only one item is associated to a genre that is not in the profile. <ul style="list-style-type: none"> • Only three items are not within the results of Test I (IOAP Avg. Diss.), which are [4],[129],[315]. First three items are the same as Test I. • Covers all relevant associations. • Only one item is associated to a genre that is not in the profile. • Missing only one genre (<i>i.e.</i>, Children's) to cover all genres in the user profile. <ul style="list-style-type: none"> • Missing only one genre (<i>i.e.</i>, Romance), to cover all genres in the user profile. • Six items are associated to genres that are not in the user profile. • Doesn't keep as much of the relevant genre associations.
Genre	Frequency																																		
Action	15																																		
Adventure	9																																		
Sci-Fi	9																																		
Drama	8																																		
Thriller	8																																		
Comedy	5																																		
Romance	5																																		
Crime	4																																		
War	4																																		
Children's	1																																		
Genre Pair	Frequency																																		
Action, Adventure	7																																		
Action, Romance	4																																		
Thriller, Action	6																																		
Sci-Fi, Action	6																																		

	<p>IV KNN IOAP Min. Diss.</p>	<ol style="list-style-type: none"> [332] Crime/Drama/Thriller [201] Action/Adventure/Comedy/Horror [89] Film-Noir/Sci-Fi [286] Drama/Romance/War [510] Action/Drama/Western [334] Action/Crime/Mystery [304] Adventure/Children's [408] Animation/Comedy/Thriller [347] Comedy/Drama [355] Adventure/Sci-Fi/Thriller 	<ul style="list-style-type: none"> Missing only one genre (<i>i.e.</i>, Romance), to cover all genres in the user profile. Only presents four items that are not within the results of Test III (KNN IOAP Avg. Diss.), which are [286],[334],[347][355]. First two items are the same as Test III. Doesn't keep as much of the relevant genre associations. Two items are associated to genres that are not in the user profile. 																																												
<p>High Heterogeneity - User 96 Size of Profile: 55</p> <p>Genre appearance frequency:</p> <table border="1" data-bbox="240 931 523 1756"> <thead> <tr> <th>Genre</th> <th>Frequency</th> </tr> </thead> <tbody> <tr><td>Drama</td><td>17</td></tr> <tr><td>Action</td><td>15</td></tr> <tr><td>Comedy</td><td>15</td></tr> <tr><td>Thriller</td><td>15</td></tr> <tr><td>Romance</td><td>11</td></tr> <tr><td>Sci-Fi</td><td>11</td></tr> <tr><td>Crime</td><td>7</td></tr> <tr><td>War</td><td>7</td></tr> <tr><td>Adventure</td><td>5</td></tr> <tr><td>Horror</td><td>5</td></tr> <tr><td>Children's</td><td>4</td></tr> <tr><td>Film-Noir</td><td>4</td></tr> <tr><td>Mystery</td><td>3</td></tr> <tr><td>Documentary</td><td>2</td></tr> <tr><td>Animation</td><td>1</td></tr> <tr><td>Fantasy</td><td>1</td></tr> <tr><td>Musical</td><td>1</td></tr> <tr><td>Western</td><td>1</td></tr> </tbody> </table> <p>Relevant genre associations:</p> <table border="1" data-bbox="240 1839 523 1993"> <thead> <tr> <th>Genre Pair</th> <th>Frequency</th> </tr> </thead> <tbody> <tr> <td>Thriller, Action</td> <td>7</td> </tr> <tr> <td>Romance,</td> <td>7</td> </tr> </tbody> </table>	Genre	Frequency	Drama	17	Action	15	Comedy	15	Thriller	15	Romance	11	Sci-Fi	11	Crime	7	War	7	Adventure	5	Horror	5	Children's	4	Film-Noir	4	Mystery	3	Documentary	2	Animation	1	Fantasy	1	Musical	1	Western	1	Genre Pair	Frequency	Thriller, Action	7	Romance,	7	<p>I IOAP Avg. Diss.</p> <p>II IOAP Min. Diss.</p> <p>III KNN IOAP Avg. Diss.</p>	<ol style="list-style-type: none"> [315] Drama/Thriller [172] Action/ Adventure/ Drama/ Romance/ Sci-Fi/ War [298]Action/Sci-Fi/Thriller [481] Comedy/Drama [402] Comedy/Romance/Thriller [504] Crime/Drama [651] Action/Drama/War [855] Action/ Drama/ Mystery/ Romance/ Thriller [523] Comedy/Drama [207] Action/Drama/Romance <ol style="list-style-type: none"> [315] Drama/Thriller [172] Action/ Adventure/ Drama/ Romance/ Sci-Fi/ War [298] Action/Sci-Fi/Thriller [402] Comedy/Romance/Thriller [481] Comedy/Drama [855] Action/ Drama/ Mystery/ Romance/ Thriller [504] Crime/Drama [641] Drama/War [531] Drama/Romance [510]Action/Drama/Western <ol style="list-style-type: none"> [1009] Drama [298] Action/Sci-Fi/Thriller [302] Crime/ Film-Noir/ Mystery/ Thriller [97] Adventure/Drama/Western [161] Action/Romance [1121]Drama/Musical [12] Crime/Thriller 	<ul style="list-style-type: none"> First item has the most frequent genre in the user profile. Thriller is also a frequent genre and the association "Drama, Thriller" appears 3 times in the user profile. The second item has two important associations and the most frequent genre. Eight genres from the user profile are missing, but have low frequency. Results cover all the important genre associations. <ul style="list-style-type: none"> First seven items are also results found in Test I (IOAP Avg. Diss.). The first three items are the same as Test I. Results cover all the important genre associations. Seven genres from the user profile are missing but have low frequency. <ul style="list-style-type: none"> Missing six genres, to cover all genres in the user profile. Missing the genre Comedy which has high frequency. Doesn't keep as much of the relevant genre associations.
Genre	Frequency																																														
Drama	17																																														
Action	15																																														
Comedy	15																																														
Thriller	15																																														
Romance	11																																														
Sci-Fi	11																																														
Crime	7																																														
War	7																																														
Adventure	5																																														
Horror	5																																														
Children's	4																																														
Film-Noir	4																																														
Mystery	3																																														
Documentary	2																																														
Animation	1																																														
Fantasy	1																																														
Musical	1																																														
Western	1																																														
Genre Pair	Frequency																																														
Thriller, Action	7																																														
Romance,	7																																														

Comedy			8. [641] Drama /War	
Sci-fi, Action	6		9. [250] Action/Sci-Fi 10. [657] Film-Noir/Thriller	
Crime, Drama	5	IV	1. [1009] Drama	<ul style="list-style-type: none"> • Missing six low frequency genres to cover all genres in the user profile. • There is an overlap of only four items with respect to results of Test III (KNN IOAP Avg. Diss.), overlapping items are: [1009][298][302][97]. The first two items are the same as Test III. • Only missing one relevant genre association.
Adventure, Action	4	KNN	2. [298] Action/Sci-Fi/Thriller	
		IOAP	3. [482] Comedy/Crime	
		Min. Diss.	4. [302] Crime/Film-Noir/Mystery/Thriller 5. [172] Action/ Adventure/ Drama/ Romance/ Sci-Fi/ War 6. [97] Adventure/ Drama /Western 7. [177] Action/Western 8. [402] Comedy/Romance /Thriller 9. [855] Action/ Drama/ Mystery/ Romance/ Thriller 10. [1159] Mystery/Sci-Fi	

Table 11. Pure exploitation qualitative test results

Overall, tests had positive results. As expected, outcomes for the low heterogeneity user profile are not as diverse as results presented for the high heterogeneity user profile. In terms of exploitation, this indicates that user profile heterogeneity is an influencing factor to determine the diversity of outcomes. Nevertheless, all pure exploitation configurations, independent of the user profile characteristics, attempted to cover as much as possible genres within the user profile; what is more, certain configurations were most successful at also choosing items containing the relevant genre associations. Also, as expected, it was found that as user profile heterogeneity grows it is harder for results to cover all the genres found in the user profile. To deal with this challenge, it can be observed from results —most significantly in outcomes for the high heterogeneity user profile—, that the technique prioritizes the task of covering the frequent genres and relevant genre associations first, and in this manner at least represent the most important aspects of the user profile if it is not possible to cover all the profile.

In synthesis, we deduce that the choice of *Importance of Associated Preference* for exploitation dimension offers results closest to the user profile while preserving the most the genre pair associations and also excluding unknown genres. Also, there is noteworthy overlap between results that are produced with the same exploitation dimension but different diversity dimension. In quantitative tests, we will further analyze the impact of the different diversity dimension alternatives. In the following section we will discuss test results for pure exploration tests.

5.2.2 Discussion on Pure Exploration Results

In the scenario of *pure exploration*, *XPLODIV* parameters are set to $\alpha = 0.0$ and $\beta = 0.0$. We run tests for different combinations of the possible instantiations of the exploration dimension and the diversity dimension, as shown in *Table 12*. As can be seen, the possible test combinations is larger than in the pure exploitation scenario. This is due to the dependency of the exploration dimension on a diversity dimension in order to measure the exploration value of an item, as explained in *chapter IV*.

To analyze test results, we observed if within suggested items there were genres that weren't within the user profile, *i.e.*, novel genres, and that results excluded genres already known to the user. In other words, because items can have both novel genres and known genres, we would like to only include items that have novel genres and that are not associated to any of the genres that can be found in the user profile.

		Exploration Dimension			
		User Profile Novelty (UPN)		Neighborhood Novelty (NN)	
		Average Dissimilarity (Avg.)	Minimum Dissimilarity (Min.)	Average Dissimilarity (Avg.)	Minimum Dissimilarity (Min.)
Diversity Dimension	Average Dissimilarity (Avg.)	I UPN Avg. Avg.	III UPN Min. Avg.	V NN Avg. Avg.	VII NN Min. Avg.
	Minimum Dissimilarity (Min.)	II UPN Avg. Min.	IV UPN Min. Min.	VI NN Avg. Min.	VIII NN Min. Min.

Table 12. Pure exploration test combinations

Test results are shown in *Table 13*, where the output generated by *XPLODIV* for the different test combinations is shown for the user with low heterogeneity. We also analyzed results generated for the medium heterogeneity user, nonetheless, we have not included these results in the document as observations were parallel to the user with low heterogeneity. Furthermore, we did not analyze results for the high heterogeneity user profile as this user has rated items for all genres except for the genre “unknown”. For a user that has rated all movie genres, in order to obtain novel recommendations, then new genre combinations would need to be offered which is harder to visually detect. We expect quantitative tests to offer further insight on the novelty aspect for users with high heterogeneity.

In *Table 13*, the column “*XPLODIV* Output” presents the ranked list of generated results by our approach. For each item in the results, we present the item’s associated genres and within square brackets the item’s identifier. In order to emphasize important observations within results, for each item, we strikethrough genres that are found in the user profile and that for exploration purposes are not desired. In the column “Observations I” we present our observations for each of the tests. In addition, the column “Observations II” offers a synthesis of observations found in the column “Observations I” for both the exploitation dimensions. In other words, we present a synthesis of all configurations for User Profile Novelty and separately for all configurations of Neighborhood Novelty.

From results, we observe that all configurations except for Test “VII NN Min. Avg.” and Test “VII NN Min. Avg.” attempted to cover the genres that were not found in the user profile. Although configuring the exploration dimension Neighborhood Novelty to use Minimum Dissimilarity produces unsuccessful results, the same exploration dimension can be used successfully using Average Dissimilarity. We infer that by using Minimum Dissimilarity it is harder for an item to be diverse (*i.e.*, if it is similar to at least one item it is deemed to be similar to the set as a whole) and because we are unifying user profiles from nearest neighbors and in addition candidate items are directly obtained from these same profiles, the technique will find the same item it’s evaluating in the unified nearest neighbor profile and immediately determine it is not novel. The only items that are not within candidate items are those that are in the user

profile, and therefore this is why this particular configuration is actually producing results that are more exploitative.

Overall, it follows from observations that the choice of *User Profile Novelty* offers results with more genres that are novel for all tests, and is also the most successful at excluding genres found in the user profile. Nevertheless, *Neighborhood Novelty* using *Average Dissimilarity* also produces successful results. Different diversity methods used for the exploration dimension offered outputs that were similar, but for us, the most positive results were obtained by the combination of *User Profile Novelty* and *Minimum Dissimilarity*. This test did not include any genres from the user profile and covered as many as possible novel genres. It is seen that all results for *User Profile Novelty* offered the same first seven results. Possible differences could be due more to the diversity dimension used to verify the diversity of results than the diversity dimension used to measure the aspect of exploration. In quantitative tests, we expect to further analyze the impact over results of using different diversity dimension selections to evaluate the diversity of selected items.

User Profile	Test	XPLoDIV Output	Observations I	Observations II														
Low Heterogeneity - User 914 Size of Profile: 23 items Genre appearance frequency: <table border="1" data-bbox="236 992 504 1294"> <thead> <tr> <th>Genre</th> <th>Frequency</th> </tr> </thead> <tbody> <tr> <td>Romance</td> <td>21</td> </tr> <tr> <td>Comedy</td> <td>13</td> </tr> <tr> <td>Drama</td> <td>12</td> </tr> <tr> <td>Musical</td> <td>2</td> </tr> <tr> <td>Action</td> <td>1</td> </tr> <tr> <td>Thriller</td> <td>1</td> </tr> </tbody> </table> Genres not in the user profile or novel genres: <ul style="list-style-type: none"> • Adventure • Animation • Children's • Crime • Documentary • Fantasy • Film-Noir • Horror • Mystery • Sci-Fi • War • Western • Unknown Note: Candidate items do not have items of the genres Fantasy or Unknown.	Genre	Frequency	Romance	21	Comedy	13	Drama	12	Musical	2	Action	1	Thriller	1	I	1. [200] Horror 2. [589] Western Avg. 3. [525] Film-Noir/Mystery Avg. 4. [48] Documentary 5. [520] Adventure/War 6. [179] Sci-Fi 7. [969] Animation/Children's 8. [156] Crime/Thriller 9. [488] Film-Noir 10. [198] Thriller	<ul style="list-style-type: none"> • Only two items have genres that can be found in the user profile. The genre Thriller has low frequency. • Covers all available novel genres. 	<ul style="list-style-type: none"> • First seven items are the same for all results. • The only genre left to cover is the genre Crime, considering that Fantasy and Unknown are not offered by candidate items.
	Genre	Frequency																
	Romance	21																
Comedy	13																	
Drama	12																	
Musical	2																	
Action	1																	
Thriller	1																	
II	1. [200] Horror 2. [589] Western Avg. 3. [525] Film-Noir/Mystery Min. 4. [48] Documentary 5. [520] Adventure/War 6. [179] Sci-Fi 7. [969] Animation/Children's 8. [156] Crime/Thriller 9. [154] Comedy 10. [165] Drama	<ul style="list-style-type: none"> • Three items have genres that can be found in the user profile. Both the genres Drama and Comedy have high frequency. • Covers all available novel genres. • Only last two items are different than Test I. The order of the first overlapping items is preserved. 	<ul style="list-style-type: none"> • We observe from candidate items that all items that contain the genre Crime also have a genre from the user profile, specifically associated to one of the following: Drama, Thriller and Comedy. • Test I, II and IV choose an item that covers Crime along with Thriller, which has the lowest frequency among the genres associated with Crime. 															
III	1. [200] Horror 2. [589] Western Min. 3. [525] Film-Noir/Mystery Avg. 4. [48] Documentary 5. [520] Adventure/War 6. [179] Sci-Fi 7. [969] Animation/Children's	<ul style="list-style-type: none"> • None of the items have genres that are in the user profile. • Does not cover the genre Crime from novel genres. • Only last three items are different than 	<ul style="list-style-type: none"> • Test III prefers to repeat genres from novel genres than include genres that 															

		8. [488] Film-Noir 9. [474] Sci-Fi/War 10. [675] Horror	Test II. The order of the first overlapping items is preserved.	are related to the user profile. • Test I, after covering Crime, has a less strong diversity influence and repeats a novel genre and also chooses to cover a genre that is not frequent in the user profile, <i>i.e.</i> , Thriller.
	IV UPN Min. Min.	1. [200] Horror 2. [589] Western 3. [525] Film-Noir/Mystery 4. [48] Documentary 5. [520] Adventure/War 6. [179] Sci-Fi 7. [969] Animation/Children's 8. [156] Crime/Thriller 9. [135] Drama/Mystery/Sci-Fi/Thriller 10. [154] Comedy	• Three items have genres that can be found in the user profile. Both the genres Drama and Comedy have high frequency. • Covers all available novel genres. • Only the last three items are different than Test III. The order of the first overlapping items is preserved.	• Test II and IV (diversity dimension as Minimum Dissimilarity), after covering Crime, prioritize diversity by choosing genres that are frequent in the user profile rather than repeat a novel genre.
	V NN Avg. Avg.	1. [488] Film-Noir 2. [589] Western 3. [48] Documentary 4. [179] Sci-Fi 5. [200] Horror 6. [969] Animation/Children's 7. [520] Adventure/War 8. [705] Musical/Romance 9. [479] Mystery/Thriller 10. [525] Film-Noir/Mystery	• Only two items have genres that can be found in the user profile. The genre Romance is the most frequently rated in the user profile. • Does not cover the genre Crime from novel genres.	• Configuring Neighborhood Novelty with Minimum Dissimilarity produces poor results. • By configuring Neighborhood Novelty with Average Dissimilarity results obtained with different diversity dimensions are the same with the exception of the last item.
	VI NN Avg. Min.	1. [488] Film-Noir 2. [589] Western 3. [48] Documentary 4. [179] Sci-Fi 5. [200] Horror 6. [969] Animation/Children's 7. [520] Adventure/War 8. [705] Musical/Romance 9. [479] Mystery/Thriller 10. [482] Comedy/Crime	• Three items have genres that can be found in the user profile. Both the genres Romance and Comedy have high frequency. • Covers all available novel genres. • Only the last item is different than Test V. The order of the first overlapping items is preserved.	

	VII NN Min. Avg.	<ol style="list-style-type: none"> 1. [28] Action/Drama/Thriller 2. [208] Comedy/Horror 3. [408] Animation/ Comedy/ Thriller 4. [200] Horror 5. [191] Drama/Mystery 6. [1197] Comedy/Drama 7. [165] Drama 8. [923] Drama 9. [61] Drama 10. [345] Comedy/Drama 	<ul style="list-style-type: none"> • Only one item does not have genres that are in the user profile. • Only covers three of the novel genres. 	
	VIII NN Min. Min.	<ol style="list-style-type: none"> 1. [28] Action/Drama/Thriller 2. [208] Comedy/Horror 3. [408] Animation/ Comedy/ Thriller 4. [200] Horror 5. [191] Drama/Mystery 6. [1197] Comedy/Drama 7. [165] Drama 8. [923] Drama 9. [61] Drama 10. [345] Comedy/Drama 	<ul style="list-style-type: none"> • Results are exactly the same as Test VII. 	

Table 13. Pure exploration qualitative test results

In this section, we have presented qualitative test results and detected which exploration and exploitation dimensions overall tend to perform the best. In the following section we will discuss results from quantitative tests.

5.3 Quantitative Offline Experiment

In this section, we analyze the results from quantitative tests. In the first place, we describe the experiment set up in *section 5.3.1*. Next, we propose a Diversity-Aware Evaluation Framework that structures metrics used to compare results of diversification techniques for *RecSys*. Lastly, using the method and selected metrics from the proposed framework, we analyze and compare results in *section 5.3.3* for different configurations of *XPLODIV* with respect to both the selected baselines and start-of-the-art techniques, which are: No Diversity, Random Diversity and *MMR*. In general, with quantitative tests we want to prove the veracity of both *Hypothesis I* and *Hypothesis II*.

5.3.1 Experiment Set Up

Given findings from qualitative tests described in *section 5.2*, for the purpose of our quantitative experiment, we configure *XPLODIV* to use as an exploitation dimension the component *ImportanceOfAssociatedPreference* and as an exploration dimension the component *UserProfileNovelty* using *MinimumDissimilarity* (reference *Table 9*, to view implementation details for used dimensions). As for the diversity dimension, we ran tests for both possible instantiations, *i.e.*, *AverageDissimilarity* and *MinimumDissimilarity*, to observe the impact different methods of measuring diversity have over results.

For all experiments we used $k=15$, thus all diversification techniques are configured to generate an ordered set of diversified items of size fifteen. We choose this size for the reason that, commonly, recommendation lists have between ten to twenty items, thus fifteen offers an intermediate size that is not too small but not too big either.

We ran experiments for different values of the *XPLODIV* control parameters as follows:

- (a) *Relevance Bias*: $\alpha = 0.8, \beta = 0.5$.
- (b) *Exploitation Bias*: $\alpha = 0.2, \beta = 0.7$.
- (c) *Exploration Bias*: $\alpha = 0.2, \beta = 0.3$.
- (d) *No Bias*: $\alpha = 0.5, \beta = 0.5$.
- (e) *Pure Exploitation*: $\alpha = 0.0, \beta = 1.0$.
- (f) *Pure Exploration*: $\alpha = 0.0, \beta = 0.0$.

We expect to prove the tunability of *XPLODIV* with results from bias configurations, which should perform in harmony to the discussed trade-offs, *e.g.*, diversity bias configurations (*i.e.*, all configurations except for relevance bias) would present an increase in diversity but a decrease on relevance, exploitation bias approaches would have low exploration values, among others.

In the following section, we will present the Diversity-Aware Evaluation Framework proposed to evaluate diversification techniques for *RecSys*.

5.3.2 Diversity-Aware Evaluation Framework

In *section 3.3*, we analyzed metrics proposed by related work that evaluate *RecSys* diversification results. We find that individually these metrics can measure different aspects of results but it is very difficult to devise an integrated metric that can offer insight of all aspects at the same time. This is why we propose a Diversity-Aware Evaluation Framework to structure metrics and identify the aspects that should be taken in to account when evaluating *RecSys* within the context of diversity. In our framework, we establish different evaluation perspectives that individually offer a view on the value of an aspect related to a recommendation list. Perspectives should be first analyzed individually and then in comparison to each other in order to obtain an integrated view of the quality of results.

We define four core perspectives which are:

- (a) Relevance Perspective
- (b) Diversity Perspective
- (c) Exploitation Perspective
- (d) Exploration Perspective

As a supplementing view, we propose a Statistical Perspective to augment insight on the core perspectives. The framework can also be extended to address other perspectives, such as an Accuracy Perspective, comprising well-known metrics such as Root Mean Squared Error.

In *Figure 11*, we present a class diagram with the proposed evaluation perspectives and possible metrics for each.

We will further discuss both perspectives and metrics on the remainder of this section. In addition, in this section we identify the metrics that are used to analyze quantitative test results. Thorough discussion on evaluation results for each of the perspectives is presented in *section 5.3.3*.

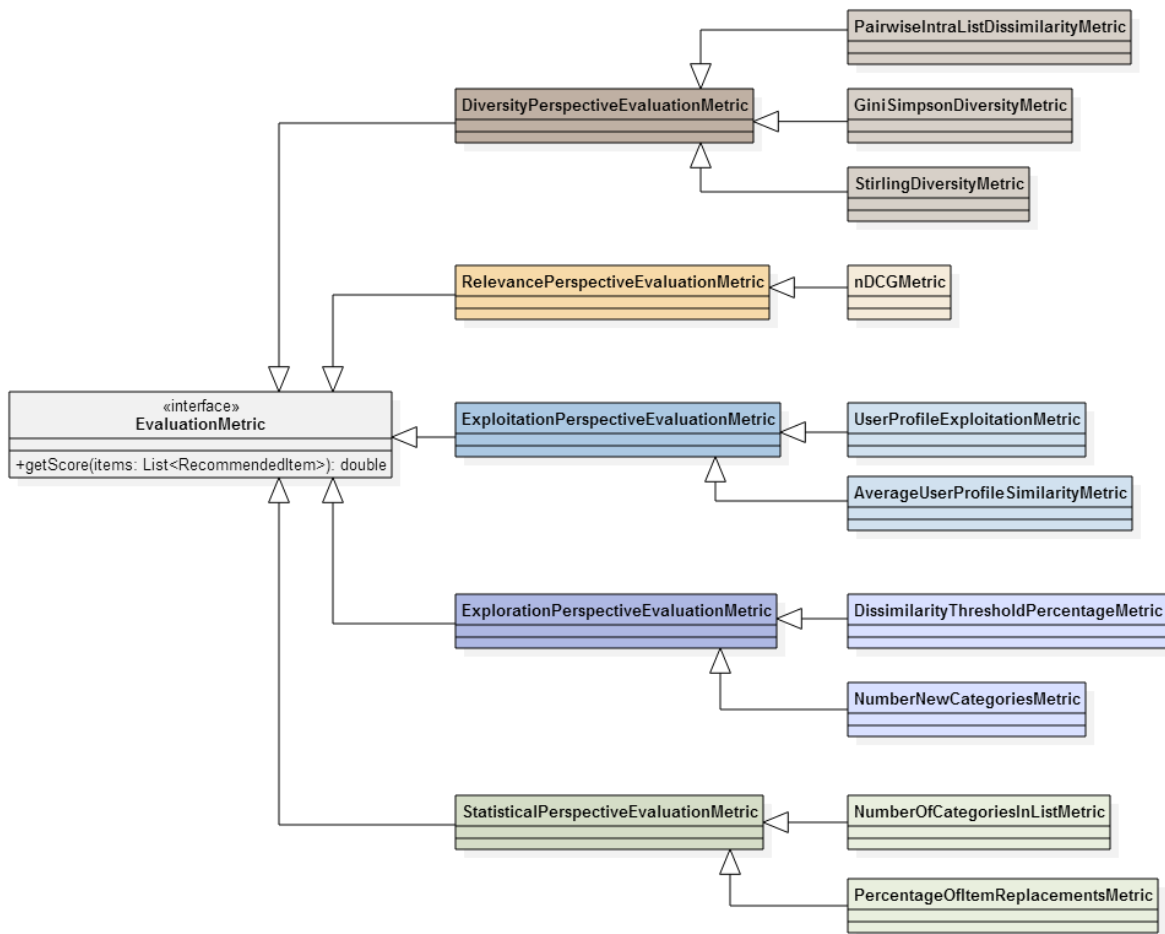


Figure 11. Diversity-Aware Evaluation Framework

Diversity Perspective

The Diversity Perspective, measures the diversity offered by a recommendation list. In both *section 2.1* and *section 3.3*, we have identified a number of metrics to measure the diversity found among elements in a set. From reviewed measures, we choose to highlight three metrics within our framework: (i) *Pairwise Intra-List Dissimilarity metric (PILD)*: specified in *Equation 2*, (ii) *Gini-Simpson Diversity index (Gini)*: specified as the complement of the Simpson measure (i.e., $Gini = 1 - Simpson$), which is specified in *Figure 3*, and (iii) *Stirling Diversity heuristic*: defined in *Equation 4*. There is an important connection between the *Gini-Simpson Diversity Index* and the *Stirling Diversity heuristic*: the “balance weighted variety” version of the Stirling heuristic, which only measures the variety and balance properties of diversity, produces equivalent results to half the value of *Gini* (view *Table 1*). In our case, given that to our knowledge information about the dissimilarity between genres themselves is not available for the MovieLens dataset (e.g., $sim(genre_{drama}, genre_{Action}) = ?$), then the “balance weighted variety” configuration of the Stirling heuristic is the only configuration we can use with the MovieLens dataset. For this reason, for quantitative tests, we will only use the *Gini* and *PILD* metrics. In future work, we plan to expand on the study of dissimilarity measures among the genres and in this way obtain results for Stirling’s diversity heuristic that actually measure the three diversity properties.

Relevance Perspective

The Relevance Perspective, evaluates a recommendation list in terms of the predicted rating or potential relevance of items. In our framework, we highlight a well-known metric which is *normalized Discounted Cumulative Gain (nDCG)* to measure the relevance of a set of recommended items. We define *nDCG* as in *Equation 49*, where the ideal *DCG* (*i.e.*, *IDCG*) is obtained from the ordered top k items from the set of candidate items \mathbb{C} , which are retrieved using the function $TopK(\mathbb{C})$. We specifically choose *nDCG*, as it is a rank-aware relevance metric that penalizes low predicted ratings at the top of the list of results.

$$nDCG(\mathbb{R}) = \frac{DCG(\mathbb{R})}{IDCG} = \frac{DCG(\mathbb{R})}{DCG(TopK(\mathbb{C}))}$$

$$DCG(\mathbb{R}) = rel(\mathbb{R}_1) + \sum_{i=2}^{|\mathbb{R}|} \frac{rel(\mathbb{R}_i)}{\log_2 i} \quad (49)$$

Where, \mathbb{R}_i is the item in position i within the ordered set \mathbb{R} .

Equation 49. Normalized Discounted Cumulative Gain (nDCG)

Exploitation Perspective

The Exploitation Perspective, evaluates how much items within results exploit known information about a target user, which is found within the user's profile. We highlight in our framework two metrics: (i) *Average User Profile Similarity metric (AUPS)*: measures the average of similarities that items within the result set \mathbb{R} have to the user profile—the inverse of the diversity of an item to a set is used to measure the similarity each item in \mathbb{R} has to the user profile—, and (ii) *User Profile Exploitation metric (UPE)*: measures the average of how well represented each item from the user profile \mathbb{U} is by items in the set \mathbb{R} (it is determined that each item in the user profile $u \in \mathbb{U}$ is represented by the item in the set \mathbb{R} that is most similar to the item u), as defined in *Equation 50*. In quantitative tests we evaluate results for both metrics. However, we find that the metric *UPE* is better suited to evaluate exploitation, as it does not penalize results if they contain explorative/novel items as the metric *AUPS*. For example, if we have a homogeneous user profile where the user has rated items of solely one genre then in results the user profile might be fully exploited by one item associated to the genre the user has rated in the past. This means that the remainder of the result list could be composed of explorative items which by definition are those furthest from the user profile. If we use the *AUPS* metric to evaluate this particular case, the result set would obtain a low exploitation value even though the user profile has already been represented by one item. In this particular case, the metric finds that items as a whole all far away from the user profile. Conversely, with the *UPE* metric, the result set would obtain a high exploitation value as this metric can consider the possibility that many items from the user profile can be represented by only one item from the result set (*i.e.*, by the item obtained from $\max_{r \in \mathbb{R}} sim(r, u)$). Nevertheless, we observe both metrics as they both offer useful insight on the behavior of results.

$$UserProfileExploitation(\mathbb{R}, \mathbb{U}) = \frac{1}{|\mathbb{U}|} \cdot \sum_{u \in \mathbb{U}} \max_{r \in \mathbb{R}} sim(r, u) \quad (50)$$

Equation 50. User Profile Exploitation metric

Exploration Perspective

The Exploration Perspective measures the amount of novelty that can be found within the list of item recommendations. To measure recommendation list novelty we highlight two metrics: (i) *Number of New Categories metric*: in the case of the MovieLens dataset, this metric counts the number of unique genres that can be found within the result set that cannot be found within the user profile, and (ii) *Dissimilarity Threshold Percentage metric (DTP)*: percentage of novel items within the set of results \mathbb{R} , where novel items are those that have a dissimilarity from the target user profile \mathbb{U} larger or equal to the threshold τ , as defined in Equation 51. In Equation 51, $d(r, u)$ is the distance of item r to item u which we measure as the inverse of the similarity between the items (i.e., $1 - sim(r, u)$). As has been mentioned, to measure the similarity between items we used the Jaccard similarity coefficient between movie genres. Lastly, through experimental observations we found that for the MovieLens dataset the value $\tau = 0.9$ was sufficiently large to observe the exploration aspect and to omit more exploitative items. We explain in detail how the value for the threshold τ was obtained in section 5.3.3.

$$DTP(\mathbb{R}, \mathbb{U}, \tau) = \frac{1}{|\mathbb{R}|} \cdot \sum_{r \in \mathbb{R}} NoveltyFilter(r, \mathbb{U}, \tau)$$

$$NoveltyFilter(r, \mathbb{U}, \tau) = \begin{cases} 1, & UserProfileDissimilarity(r, \mathbb{U}) \geq \tau \\ 0, & otherwise \end{cases} \quad (51)$$

$$UserProfileDissimilarity(r, \mathbb{U}) = \frac{1}{|\mathbb{U}|} \cdot \sum_{u \in \mathbb{U}} d(r, u)$$

Equation 51. Dissimilarity Threshold Percentage metric

For our experiments, we only used the *DTP* metric, since the *Number of New Categories metric* cannot account for the fact that different genre combinations, even if the genres can be found within the user profile, could represent novelty. This is especially important for user profiles with high heterogeneity, which according to the *Number of New Categories metric* would never obtain novel results. Even so, results from this metric helped verify conclusions from the qualitative tests.

Statistical Perspective

The Statistical Perspective is meant to offer additional insight on the core perspectives. We have proposed two metrics for this perspective: (i) *Number of Categories in List metric*: in the case of the MovieLens dataset, this metric measures the number of unique genres that can be found in a result set, and (ii) *Percentage of Item Replacements metric*: evaluates the percentage of items from No Diversity that have been replaced, or cannot be found, within results that have been diversified.

The Statistical Perspective could be further augmented with measurements that help observe characteristics of the generated results sets. Additional metrics could also be added to analyze the characteristics of candidate items, such as the availability of novel items among the set of

candidate items, the similarity among the items that are novel, the similarity in general of items in the candidate set, among others. Analyzing the candidate set of items could offer valuable insight to understand results from the other perspectives.

For our experiments, we observed results from both of the proposed metrics for the Statistical Perspective.

In this section, we have proposed a Diversity-Aware Evaluation Framework that defines and structures metrics to be used to carry out quantitative tests. In the following section we will discuss results obtained from tests using the presented metrics.

5.3.3 *Data Interpretation and Discussion*

In this section, we evaluate the outputs for our approach, baseline and state-of-the-art techniques using the Diversity-Aware Evaluation Framework for *RecSys* proposed in the previous section. Before anything else, we show how we determined the value of the threshold τ for the *Dissimilarity Threshold Percentage metric (DTP)*. After this, we analyze results for the perspectives of diversity, relevance, exploitation, exploration and statistical. Finally, we discuss an integrated view of metrics to observe the trade-offs among the different aspects within results.

Results for each of the metrics from the different perspectives are shown in three graphs:

- (i) *Average Dissimilarity Graph*: presents results for bias set-ups (*i.e.*, excluding pure set-ups: Pure Exploration and Pure Exploitation) compared to baselines and state-of-the-art techniques, when using Average Dissimilarity as diversity dimension for all *XPLODIV* configurations.
- (ii) *Minimum Dissimilarity Graph*: presents results for bias set-ups compared to baselines and state-of-the-art techniques, when using Minimum Dissimilarity as diversity dimension for all *XPLODIV* configurations.
- (iii) *Tuning Graph*: displays only the pure configuration set-ups (*i.e.*, Pure Exploration and Pure Exploitation) for both diversity dimensions, compared to the No Diversity results. Each of the graphs, shows results from the different techniques for user profiles with different levels of heterogeneity, where heterogeneity levels are in the range [0,100].

On the one hand, by comparing the *Average Dissimilarity Graph* to the *Minimum Dissimilarity Graph*, we hope to observe the impact the selection of diversity dimension has over the evaluated perspective. Also, in both these graphs we can observe how the different configurations of *XPLODIV* perform compared to baselines and state-of-the-art techniques. On the other hand, with the *Tuning Graph* we want to observe how pure configurations do offer tuning control over *XPLODIV* results compared to No Diversity. With the *Tuning Graph* we can also gain insight on the impact of diversity dimension changes.

Defining the dissimilarity threshold for the Number of Dissimilar items metric

In *Figure 12*, we present a histogram of the distances each item in a list of recommendation results has to the user profile, where recommendation lists were produced by Pure Exploration and Pure Exploitation *XPLODIV* configurations.

Specifically, there are two superimposed histograms: (i) the *Pure Exploitation* histogram, which is created with distances from results from both, Pure Exploitation using Average Dissimilarity and Pure Exploitation using Minimum Dissimilarity, and (ii) the *Pure Exploration* histogram, which is likewise generated from distances obtained from Pure Exploration results using both

diversity dimensions. In brief, this diagram allows us to observe the distribution of item distances to the user profile found in generated results, and observe which range of distances are the most frequent for each configuration. Each histogram presented in *Figure 12* has 50 bins or ranges of distance values. Because the histograms have the same number of bins, and values for both are in the same range $[0,1]$, we can superimpose the histograms to compare results.

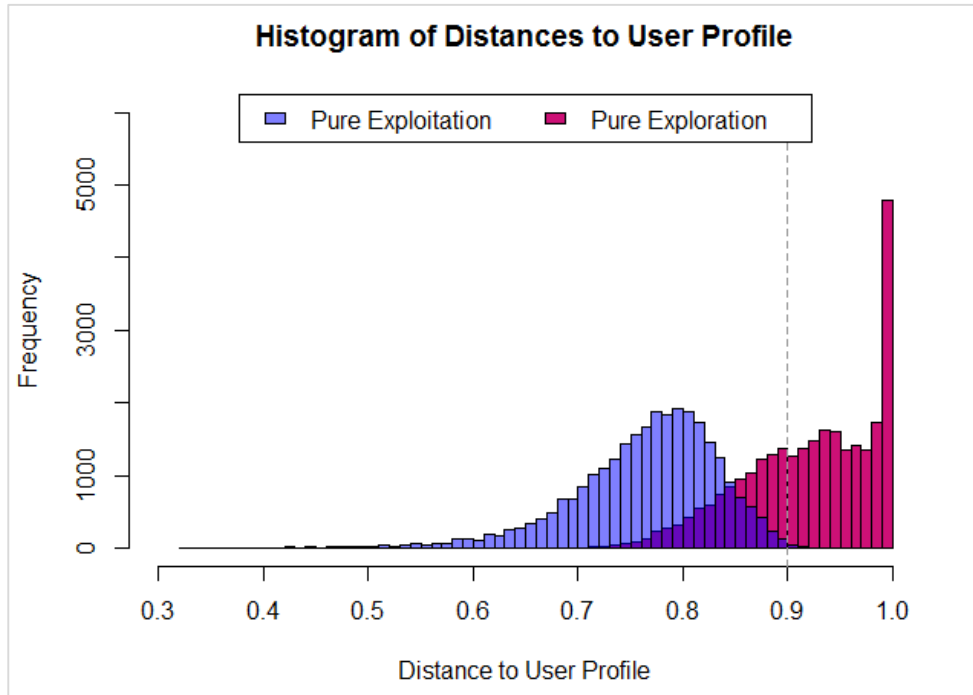


Figure 12. Histogram of Distances to User Profile

In the diagram, values closer to one are of items that are the farthest from the user profile. As can be seen, Pure Exploration items tend to be closer to one, having a high peak of items that can be placed in the last bin. Alternatively, Pure Exploitation results do not approach distances of one. However, there is a high concentration of items around distance 0.8 for Pure Exploitation items. This is normal because as user profiles grow in heterogeneity it is hard for one item to be similar to all items in the user profile.

With the histograms of pure configuration set-ups, we can further confirm the findings of qualitative tests and demonstrate that *XPLDIV* can be tuned towards more exploitative or explorative results. Moreover, in the diagram, it can be observed that the distance 0.9 can be a good threshold to determine when items start becoming explorative and have less exploitative characteristics. Specifically, we can see that to the right of the value 0.9, the frequencies for Pure Exploitation are extremely small in comparison to the rest and thus we could infer that items from this point forward have the lowest exploitative values. This is why we choose $\tau = 0.9$ as a good threshold for the for the *DPT* metric, which when used with the MovieLens 100k dataset, would offer insight on the Exploration Perspective of generated results.

In this section, we have determined how we choose the dissimilarity threshold τ for the *Dissimilarity Threshold Percentage metric*. In the following sections, we will analyze results from each of the perspectives presented in the Diversity-Aware Evaluation Framework, which are: diversity, relevance, exploitation, exploration and statistical.

Evaluation results for the Diversity Perspective

We used two metrics to analyze the Diversity Perspective: *Gini-Simpson Diversity Index (Gini)* and *Pairwise Intra-List Dissimilarity metric (PILD)*. We will refer from this point forward, to the graphs representing results, as follows: (a) *Figure 13* as *Gini* [Graph A], (b) *Figure 14* as *Gini* [Graph B], (c) *Figure 15* as *Gini* [Graph C], (d) *Figure 16* as *PILD* [Graph A], (e) *Figure 17* as *PILD* [Graph B], and (f) *Figure 18* *PILD* [Graph C].

On the one hand, from results generated by means of the *Gini* index, we observe:

- *MMR* obtains the highest diversity values using both diversity dimension configurations, *i.e.*, according to *Gini* [Graph A] and *Gini* [Graph B]. This means that *MMR*, is the technique that offers the highest diversity when assessing the properties of variety and balance.
- When using Minimum Dissimilarity, in *Gini* [Graph B], the techniques of *MMR*, Exploration Bias, Exploitation Bias, and No Bias all have similarly high diversity. It makes sense that using *XPLODIV* with the same diversity dimension as *MMR* would lead to similar characteristics within results in terms of diversity.
- In general, diversification techniques *MMR*, No Bias, Exploitation Bias and Exploration Bias; beat No Diversity, Random Diversity and Relevance Bias *XPLODIV* in terms of diversity.
- It is found, according to *Gini*, that the Exploration Bias configuration, in all graphs, produces results more diverse than Exploitation Bias.
- Observing the *Gini* [Graph C], we could infer that Minimum Dissimilarity is a better diversity dimension when evaluating for *Gini* diversity index.
- In the *Gini* [Graph C], we can see that all configurations except for Pure Exploitation using Average Dissimilarity, beat No Diversity in terms of diversity. However, we will see that with the *PILD* metric, Pure Exploitation for both diversity dimensions performs well in terms of diversity.

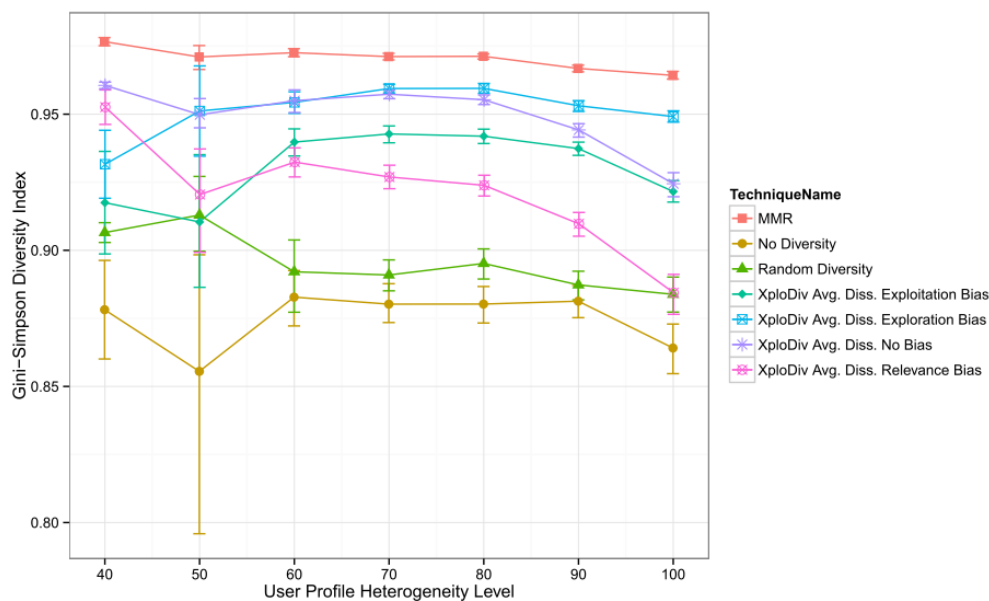


Figure 13. *Gini-Simpson Diversity Index* — [Graph A] *Average Dissimilarity Graph*

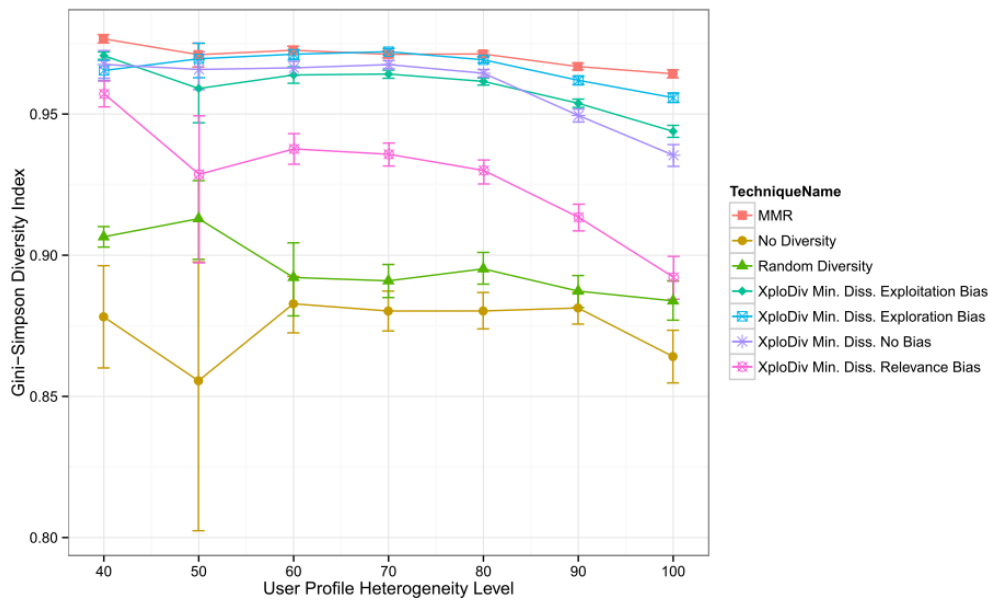


Figure 14. Gini-Simpson Diversity Index — [Graph B] Minimum Dissimilarity Graph

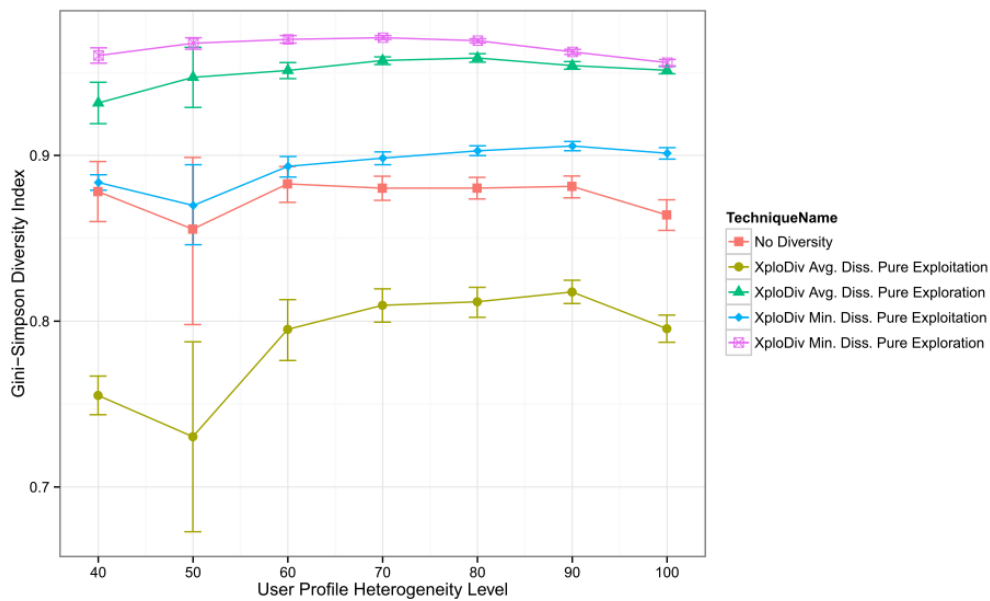


Figure 15. Gini-Simpson Diversity Index — [Graph C] Tuning Graph

On the other hand, we study results generated by means of the *PILD* metric. This metric analyzes how dissimilar items in the result set are from each other, implicitly considering not only the variety and balance properties, but also explicitly taking into account genre associations found among items. From results obtained with the *PILD* metric, we observe:

- MMR* has low diversity according to *PILD* for both diversity dimensions, *i.e.*, *PILD* [Graph A] and *PILD* [Graph B]. What is more, when the user profile has high heterogeneity, *MMR* performs the worst in terms of diversity compared to all techniques. When we analyzed the statistical metric *Number of Categories in List metric*, particularly Figure 31 and Figure 32 —which will be further discussed later on—, we found that *MMR* has the highest number of associated genres among all the techniques. This could explain why *MMR* performs better

when evaluated using *Gini*, as it maximizes variety and balance with a high number of associated genres per item. However, it could be reasoned that because items have more associated genres it is most probable that they are more similar to each other as well, which would explain the low performance of *MMR* with *PILD*.

- In general, Exploitation Bias and Exploration Bias results found in *PILD* [Graph A] and *PILD* [Graph B], present higher values of diversity as the user profile heterogeneity grows. We could infer that:
 - For Exploitation Bias, it is expected that diversity within results grows as the user profile heterogeneity grows.
 - It can be possible that Exploration Bias diversity grows with user profile heterogeneity because, as the user profile grows, it is harder for results to be explorative. This means that with higher user profile heterogeneity, Exploration Bias results might be more exploitative than results for lower user profile heterogeneity users, and therefore, grow in diversity along with the user profile heterogeneity. This idea is further confirmed when analyzing the Exploitation Perspective.
- When observing the *PILD* [Graph C], we could conclude that Minimum Dissimilarity is a better diversity dimension when evaluating for *PILD*.
- It is found, according to *PILD*, that Exploitation Bias in all graphs produces results more diverse than Exploration Bias. This is contrary to what is shown by *Gini* results.
- In general, Exploitation Bias and Pure Exploitation approaches beat No Diversity in terms of diversity. It is found that exploration approaches have slightly lower values of diversity than No Diversity. In future work, we could analyze the characteristics of candidate items to analyze if there is enough diversity among available explorative items when using *PILD* metric.

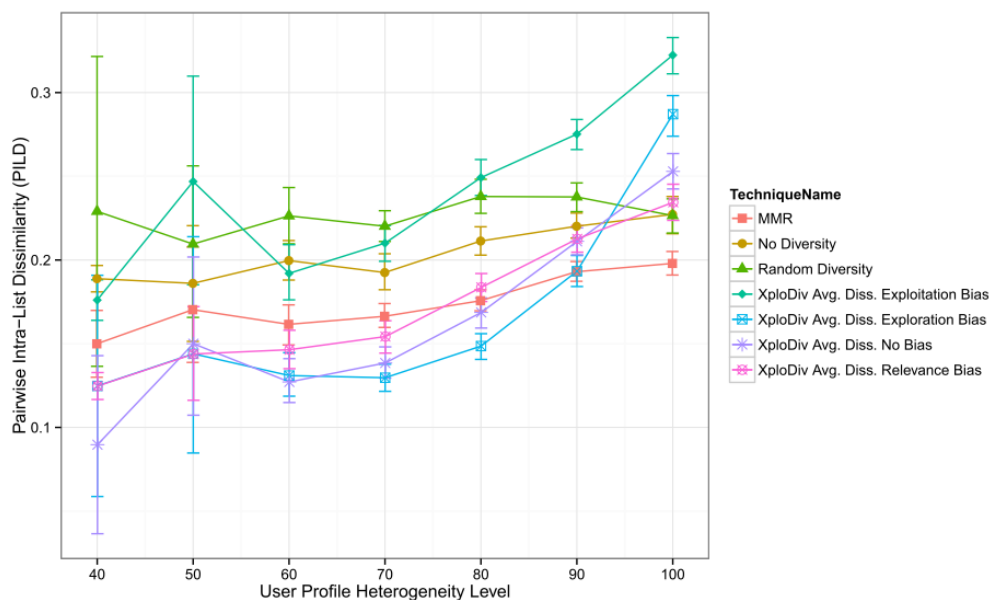


Figure 16. Pairwise Intra-List Dissimilarity Metric — [Graph A] Average Dissimilarity Graph

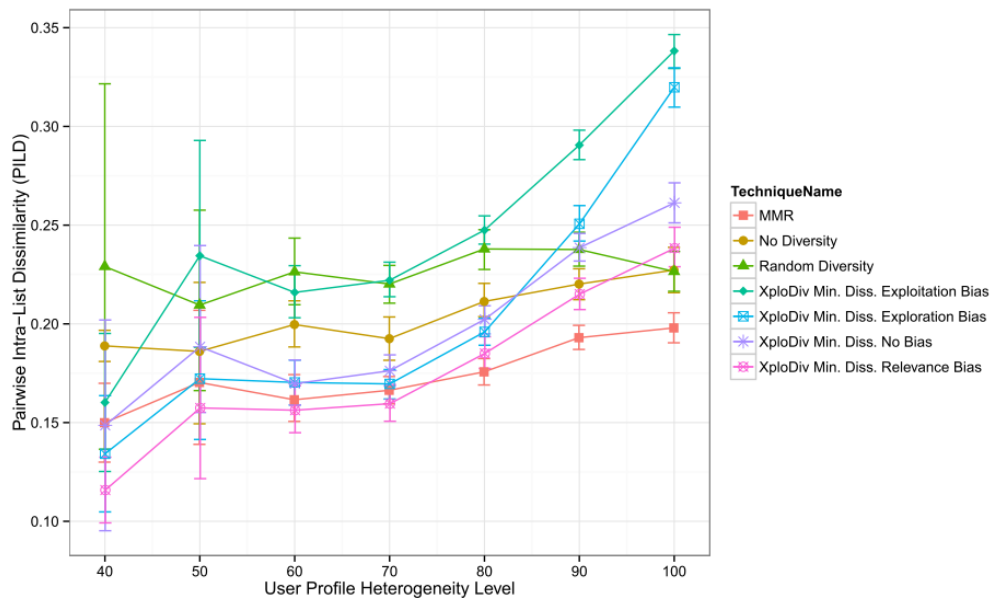


Figure 17. Pairwise Intra-List Dissimilarity Metric — [Graph B] Minimum Dissimilarity Graph

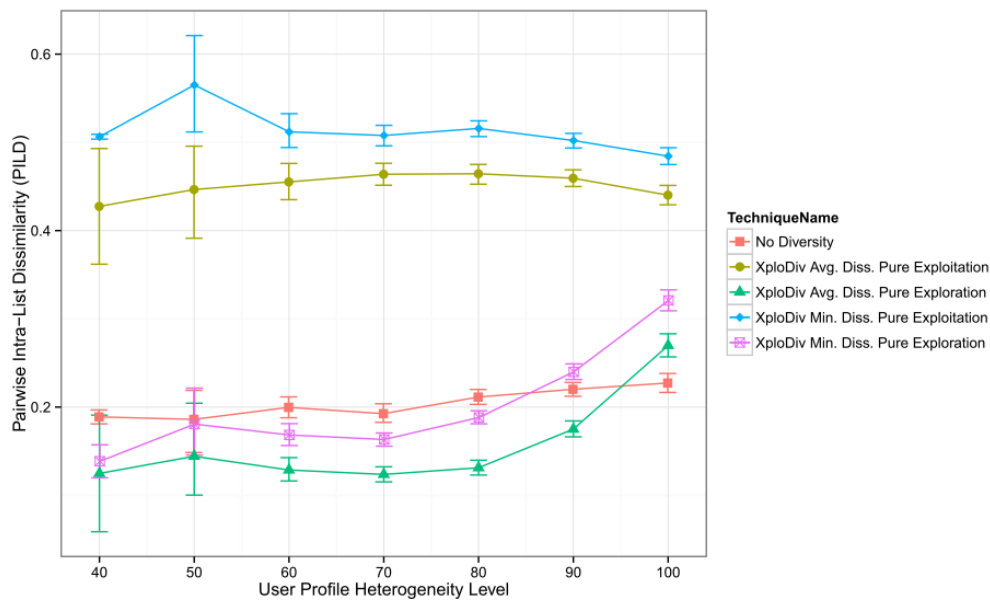


Figure 18. Pairwise Intra-List Dissimilarity Metric — [Graph C] Tuning Graph

Overall, from the Diversity Perspective we can conclude that *XPLODIV* can be tuned towards more or less diversity. In general, results for our approach were comparable or better than baselines and state-of-the-art approaches. We also find a clear influence of the user profile heterogeneity on the obtained diversity of results. However, the diversity dimension Minimum Dissimilarity, generated better results for the Diversity Perspective for both of the observed metrics.

In the following section, we will discuss results from the Relevance Perspective.

Evaluation results for the Relevance Perspective

We evaluated results using the *Normalized Discounted Cumulative Gain (nDCG)* metric to examine the Relevance Perspective. We will refer from this point forward, to the graphs representing results, as follows: (a) *Figure 19* as *nDCG* [Graph A], (b) *Figure 20* as *nDCG* [Graph B], and (c) *Figure 21* as *nDCG* [Graph C].

From results, we observe that:

- Results for No Diversity represent the ideal *nDCG*, as the metric was configured this way.
- Random Diversity produces the worst results in terms of relevance in all graphs.
- Exploration Bias results present low relevance in both *nDCG* [Graph A] and *nDCG* [Graph B]. This is expected, as novel items tend to receive lower predicted rating than exploitative items.
- In general, the No Bias configuration performs similar to *MMR*.
- The Relevance Bias configuration performs similar to No Diversity. This shows that *XPLODIV* can be tuned to be relevance bias.
- Exploitation configurations, in both *nDCG* [Graph A] and *nDCG* [Graph B], have higher relevance than exploration configurations.
- Exploitation Bias, in both *nDCG* [Graph A] and *nDCG* [Graph B], has lower relevance than No Diversity, *MMR* and No Bias. This indicates there is a sacrifice in relevance as diversity grows.
- From the *nDCG* [Graph C], we can observe that in terms of relevance the diversity dimension does not seem to have an influence over results.
- It curious to highlight that Pure Exploration results in the *nDCG* [Graph C] present higher relevance than Pure Exploitation results. Again, this could be due to the diversity *vs.* relevance trade-off and the higher diversity offered by Pure Exploitation results.
- From the *nDCG* [Graph C], we can note that as user profile heterogeneity grows, relevance diminishes.

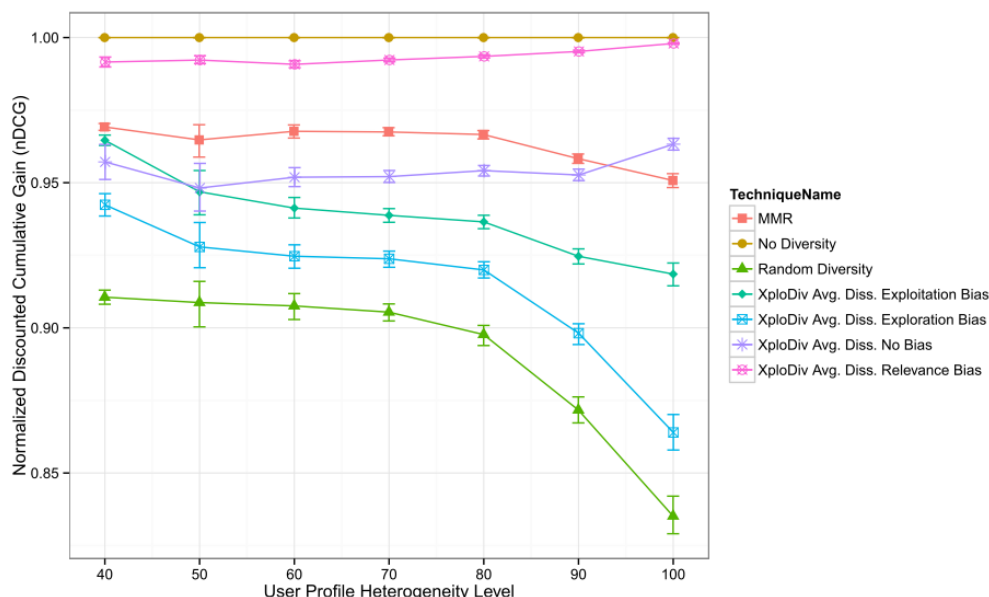


Figure 19. Normalized Discounted Cumulative Gain — [Graph A] Average Dissimilarity Graph

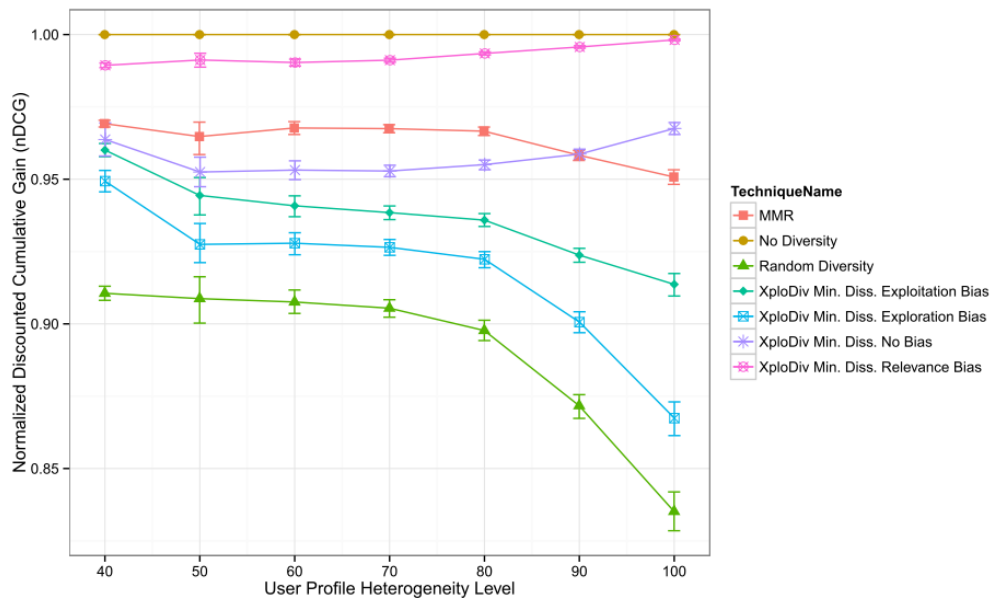


Figure 20. Normalized Discounted Cumulative Gain — [Graph B] Minimum Dissimilarity Graph

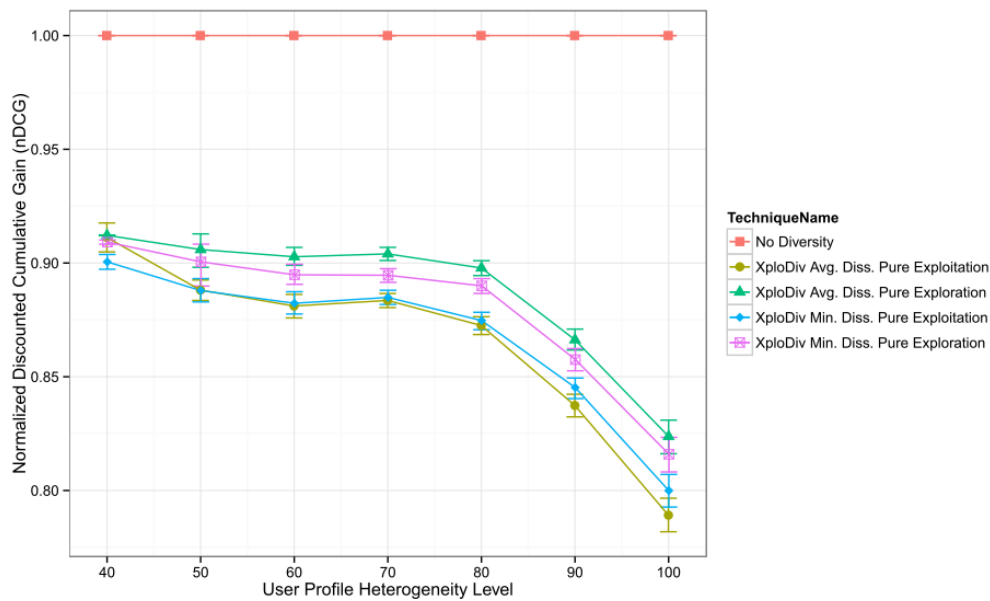


Figure 21. Normalized Discounted Cumulative Gain — [Graph C] Tuning Graph

Overall, from the Relevance Perspective we can observe that *XPLODIV* can be tuned towards results that produce higher relevance. In addition, we can see that there is a trade-off between diversity and relevance when comparing results from the Relevance Perspective to results from the Diversity Perspective. Specifically, Exploitation Bias *XPLODIV* configurations present lower relevance but higher diversity. Lastly, we observe that the Relevance Perspective does not seem to be influenced by the selection of diversity dimension.

In the following section, we will analyze results from the Exploitation Perspective.

Evaluation results for the Exploitation Perspective

We used two metrics to analyze the Exploitation Perspective: *User Profile Exploitation metric (UPE)* and *Average User Profile Similarity metric (AUPS)*. We will refer from this point forward, to the graphs representing results, as follows: (a) *Figure 22 as UPE [Graph A]*, (b) *Figure 23 as UPE [Graph B]*, (c) *Figure 24 as UPE [Graph C]*, (d) *Figure 25 as AUPS [Graph A]*, (e) *Figure 26 as AUPS [Graph B]*, and (f) *Figure 27 as AUPS [Graph C]*.

On the one hand, from results generated by means of the *UPE* metric, we observe:

- From *UPE* [Graph A] and *UPE* [Graph B], we observe that Random Diversity, No Diversity, Relevance Bias and *MMR* generate the highest exploitation value. A logical explanation could be that the amount of exploitative items in the candidate item set is much larger than explorative items, thus it is most probable for a random selection to choose an exploitative item.
- There could be a relation between the aspects of relevance and exploitation, which can be observed in the high exploitation values of No Diversity and Relevance Bias.
- It can be observed that, Exploitation Bias results for both diversity dimensions have comparable results to techniques that produce the highest exploitation values.
- Furthermore, with results from *UPE* [Graph A] and *UPE* [Graph B], we confirm the idea that Exploration Bias results become more exploitative as user profile heterogeneity grows. As expected, Exploration Bias results in general have lower exploitation value than other techniques due to the tradeoff between exploration and exploitation.
- No Bias results are found in between the Exploration Bias and Exploitation Bias results, further demonstrating that the control parameters allow to tune the exploitation vs. exploration trade-off.
- From *UPE* [Graph C], we can observe that Pure Exploitation results for both diversity dimensions are better at exploiting the user profile than No Diversity.
- From *UPE* [Graph C], we could infer that Minimum Dissimilarity is a better diversity dimension for the purpose of exploiting the user profile when evaluating with *UPE* metric. Results from *UPE* [Graph A] and *UPE* [Graph B], offer consistent observations with this interpretation.

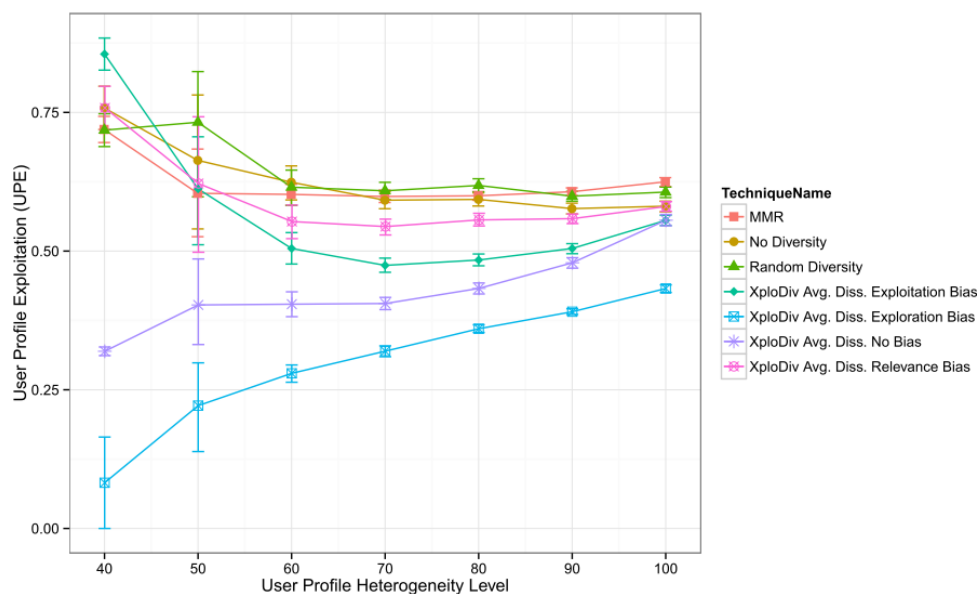


Figure 22. User Profile Exploitation Metric — [Graph A] Average Dissimilarity Graph

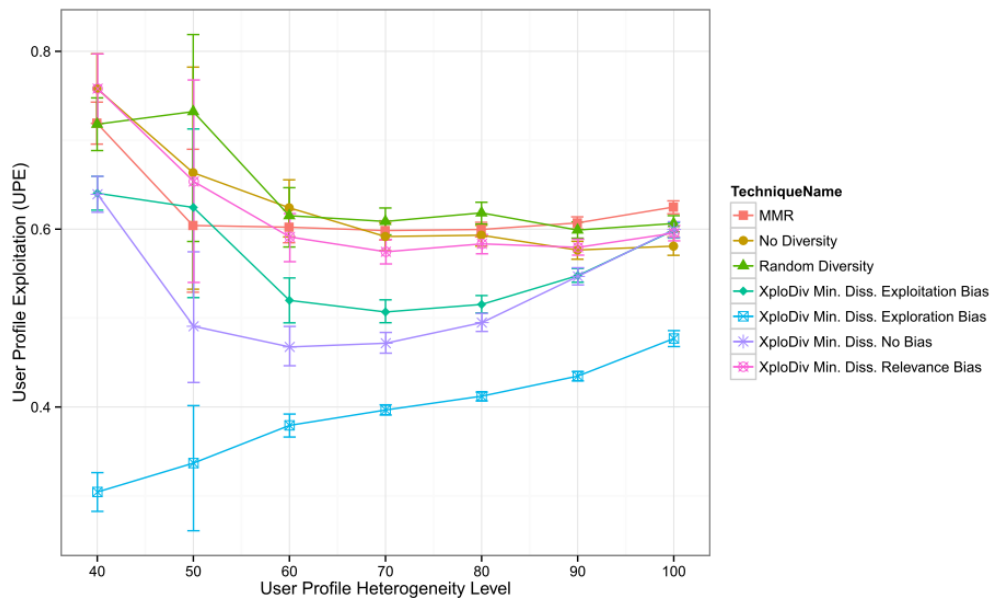


Figure 23. User Profile Exploitation Metric — [Graph B] Minimum Dissimilarity Graph

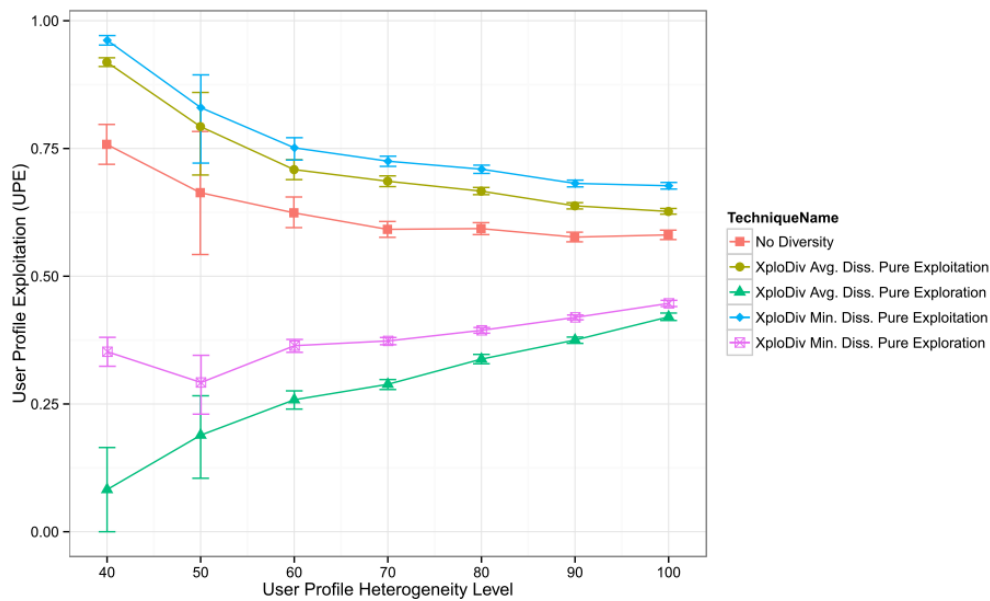


Figure 24. User Profile Exploitation Metric — [Graph C] Tuning Graph

On the other hand, from results generated by means of the *AUPS* metric, we observe:

- As found when analyzing *UPE* metric results, No Diversity and Random Diversity have high exploitation values for both diversity dimensions.
- In *AUPS* [Graph A], Exploitation Bias produces higher exploitative results than Relevance Bias and MMR. In *AUPS* [Graph B], Exploitation Bias has a higher exploitation value than *MMR* and is consistent with the Relevance Bias results.
- As expected, Exploration Bias has the lowest exploitation value, but also exploitative value grows with user profile heterogeneity. This is an important observation in favor of the Exploration Perspective as it shows that Exploration Bias results are the furthest from the user profile.

- No Bias results can be found in between Exploitation Bias and Exploration Bias results.
- No Bias results are close to *MMR* when using the same diversity metric as *MMR*, *i.e.*, Minimum Dissimilarity.
- From *AUPS* [Graph C], we can observe that Pure Exploitation results for both diversity dimensions are better at exploiting the user profile than No Diversity.
- From *AUPS* [Graph C] and *AUPS* [Graph A], we could infer that Average Dissimilarity is a better diversity dimension for exploiting the user profile when evaluating with *AUPS*. Nonetheless, from *AUPS* [Graph C] and *AUPS* [Graph B], we can observe that Minimum Dissimilarity achieves a comparable exploitation value.

Overall, from the Exploitation Perspective we can observe that *XPLODIV* can be tuned towards results that produce higher exploitation value than No Diversity and even comparable to state-of-the-art techniques. In addition, we observe a possible dependence between the aspect of relevance and the aspect of exploitation due to the high exploitation value of relevance focused techniques, such as No Diversity and Relevance Bias. Lastly, we observe that the selection of diversity dimension is not straightforward, as both Average Dissimilarity and Minimum Dissimilarity offer successful results. We could incline towards Minimum Dissimilarity, as we have argued that *UPE* could be a more appropriate exploitation metric (view *section 5.3.2*).

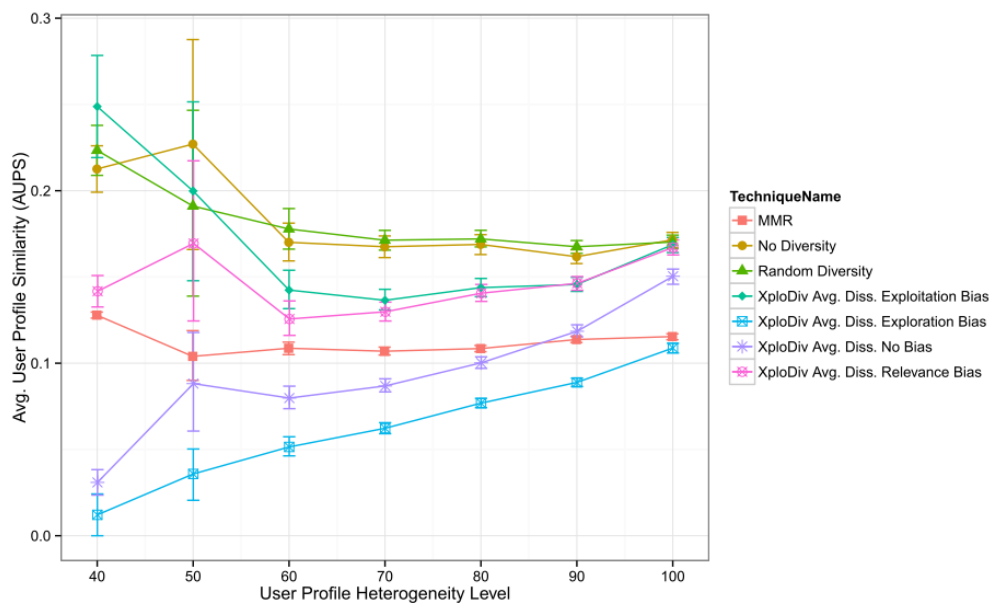


Figure 25. Average User Profile Similarity Metric — [Graph A] Average Dissimilarity Graph

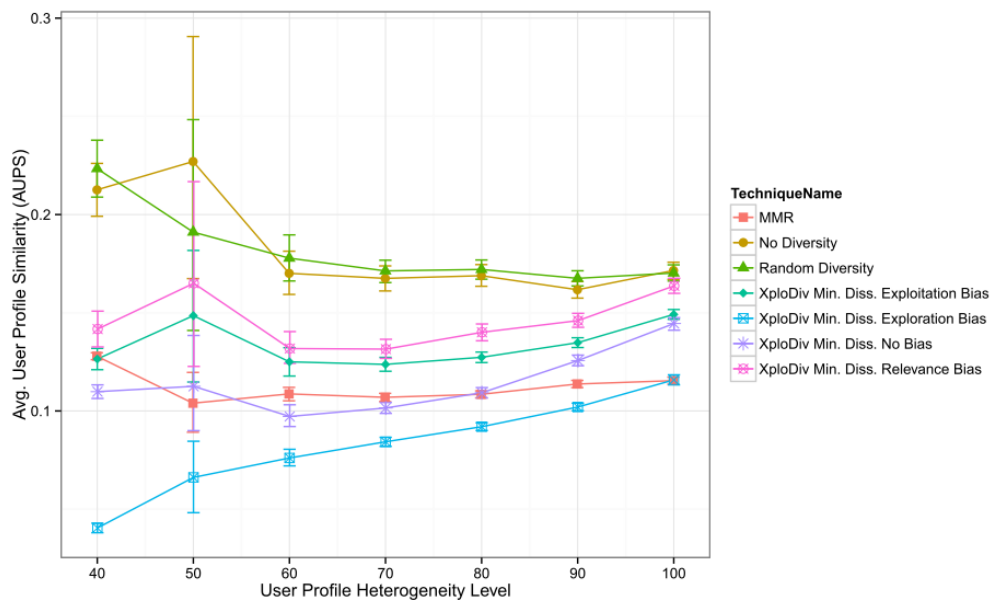


Figure 26. Average User Profile Similarity Metric — [Graph B] Minimum Dissimilarity Graph

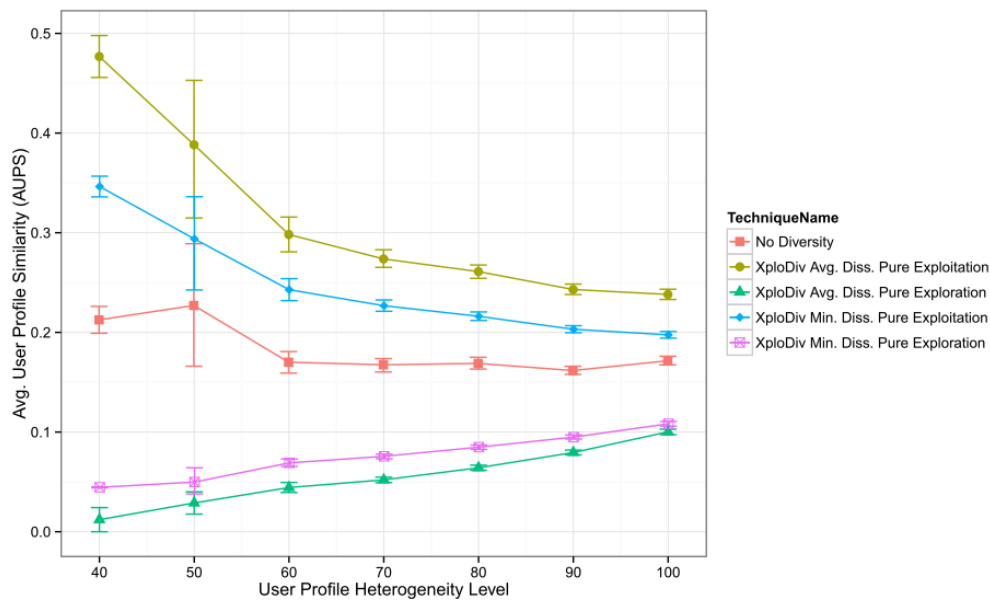


Figure 27. Average User Profile Similarity Metric — [Graph C] Tuning Graph

In the following section, we will analyze results from the Exploration Perspective.

Evaluation results for the Exploration Perspective

We evaluated results using the *Dissimilarity Threshold Percentage metric (DTP)* metric to examine the Exploration Perspective. We will refer from this point forward, to the graphs representing results, as follows: (a) Figure 28 as DTP [Graph A], (b) Figure 29 as DTP [Graph B], and (c) Figure 30 as DTP [Graph C].

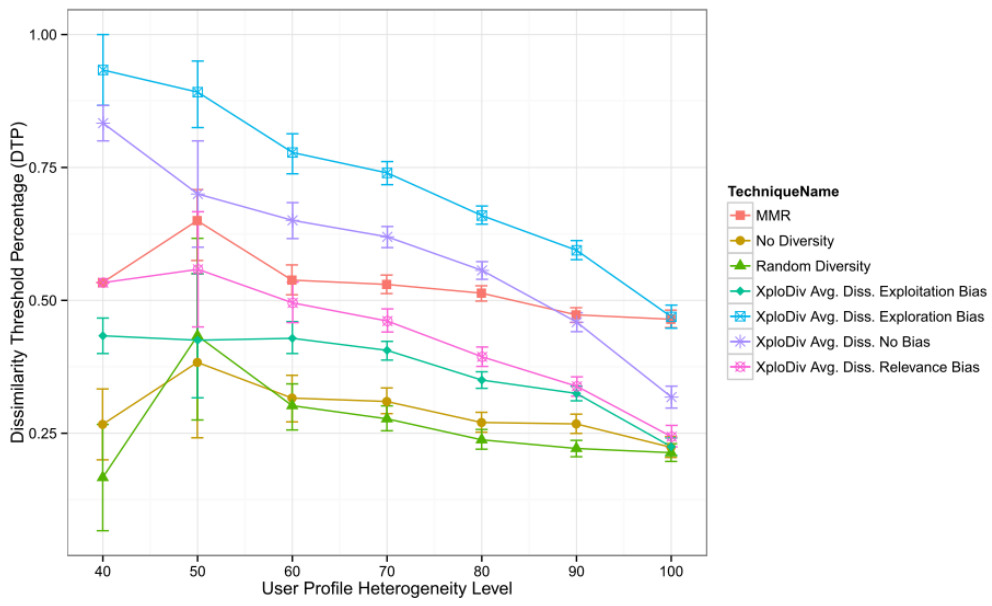


Figure 28. Dissimilarity Threshold Percentage Metric — [Graph A] Average Dissimilarity Graph

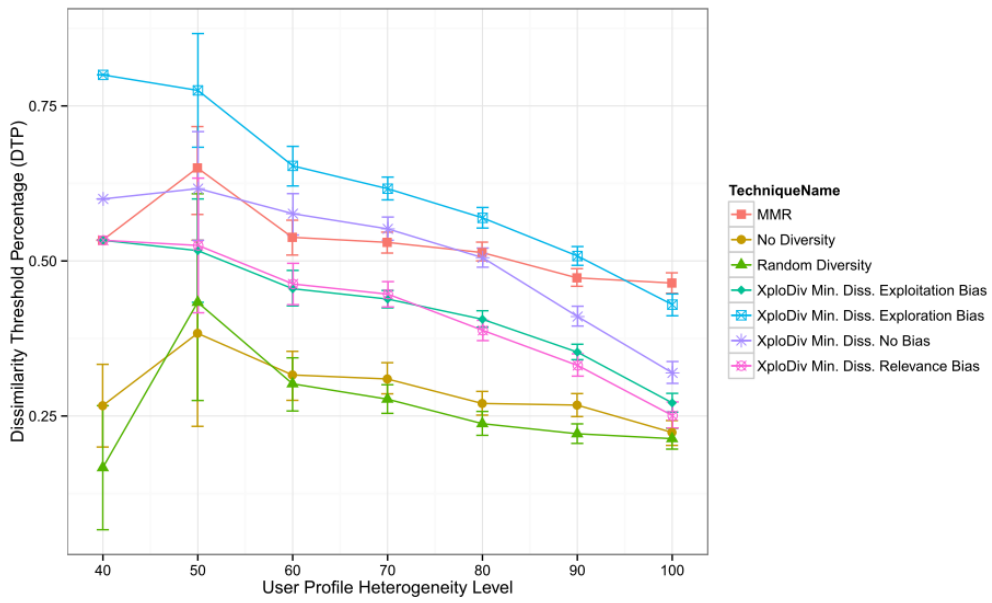


Figure 29. Dissimilarity Threshold Percentage Metric—[Graph B] Minimum Dissimilarity Graph

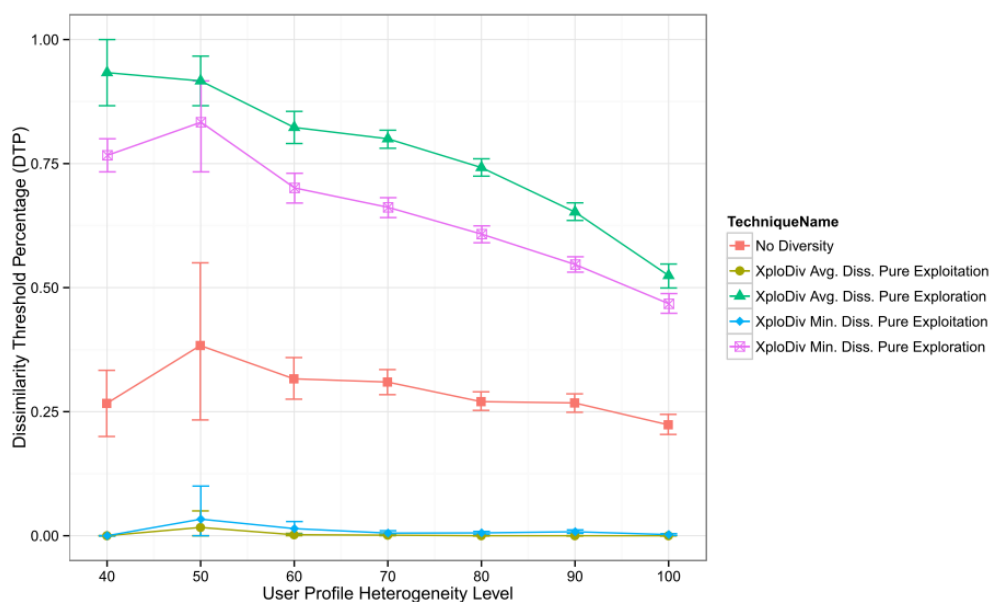


Figure 30. Dissimilarity Threshold Percentage Metric — [Graph C] Tuning Graph

From results generated by means of the *DTP* metric, we observe:

- Exploration Bias, produces the highest exploration values in all graphs.
- No Diversity and Random Diversity have the lowest exploration values, as can be viewed in both *DTP* [Graph A] and *DTP* [Graph B]. This is consistent with the exploration *vs.* exploitation, trade-off given that these two techniques have very high exploitation values.
- Exploitation Bias and Relevance Bias also have very low exploration values, which can be explained by the exploration *vs.* exploitation trade-off. However, these approaches have a larger amount of exploration value than No Diversity and Random Diversity.
- In general, as user profile heterogeneity grows, exploration values decrease. This can be observed in most techniques. The impact of user profile heterogeneity is less noticeable in *MMR*.
- From *DTP* [Graph C], we can observe that Pure Exploration always produces higher exploration value than No Diversity.
- From *DTP* [Graph C], we can observe that Pure Exploitation approaches have near zero exploration value. This confirms the idea that, if exploration is not explicitly accounted for, it is hard for a novel product to be recommended to the user.
- From *DTP* [Graph C] and *DTP* [Graph A], we could infer that Average Dissimilarity is a better diversity dimension when the goal is to explore. Nonetheless, results with Minimum Dissimilarity are also successful.

Overall, from the Exploration Perspective we can observe that *XPLODIV* can be tuned towards results that produce higher exploration value than No Diversity and state-of-the-art techniques. We can also observe the influence of the trade-off between exploration *vs.* exploitation when comparing results from both Exploration and Exploitation Perspectives. Lastly, we found that Average Dissimilarity is a better diversity dimension to generate results of a more explorative nature.

In the following section, we will analyze results from the Statistical Perspective.

Evaluation results for the Statistical Perspective

We evaluated results using the *Number of Categories in List metric (NumCategories)* and the *Percentage of Item Replacements metric (ItemsReplaced)*, to examine the Statistical Perspective. We will refer from this point forward to the graphs representing results as follows: (a) *Figure 31 as NumCategories [Graph A]*, (b) *Figure 32 as NumCategories [Graph B]*, (c) *Figure 33 as NumCategories [Graph C]*, (d) *Figure 34 as ItemsReplaced [Graph A]*, (e) *Figure 35 as ItemsReplaced [Graph B]*, and (f) *Figure 36 as ItemsReplaced [Graph C]*.

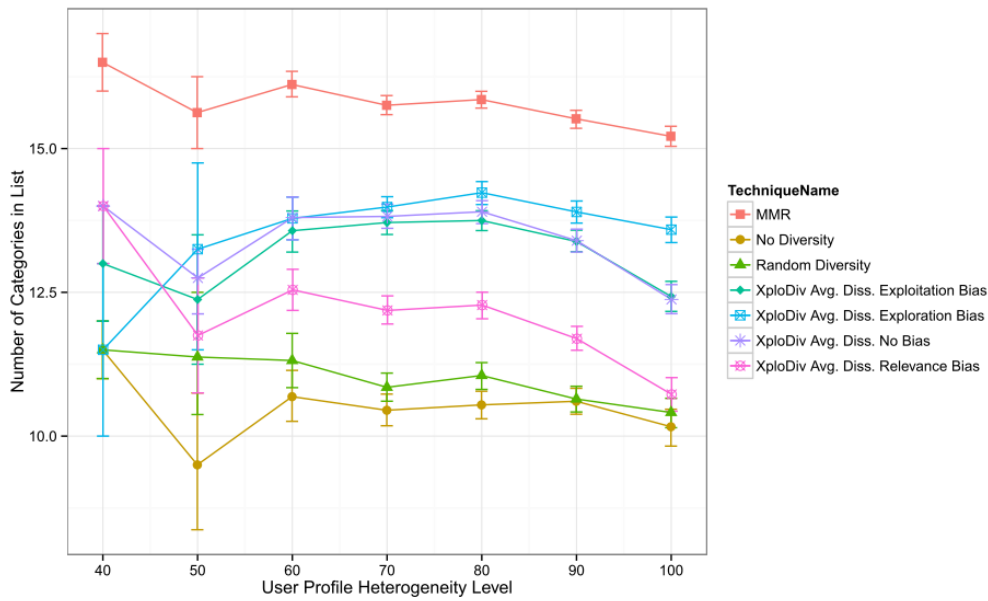


Figure 31. Number of Categories in List Metric — [Graph A] Average Dissimilarity Graph

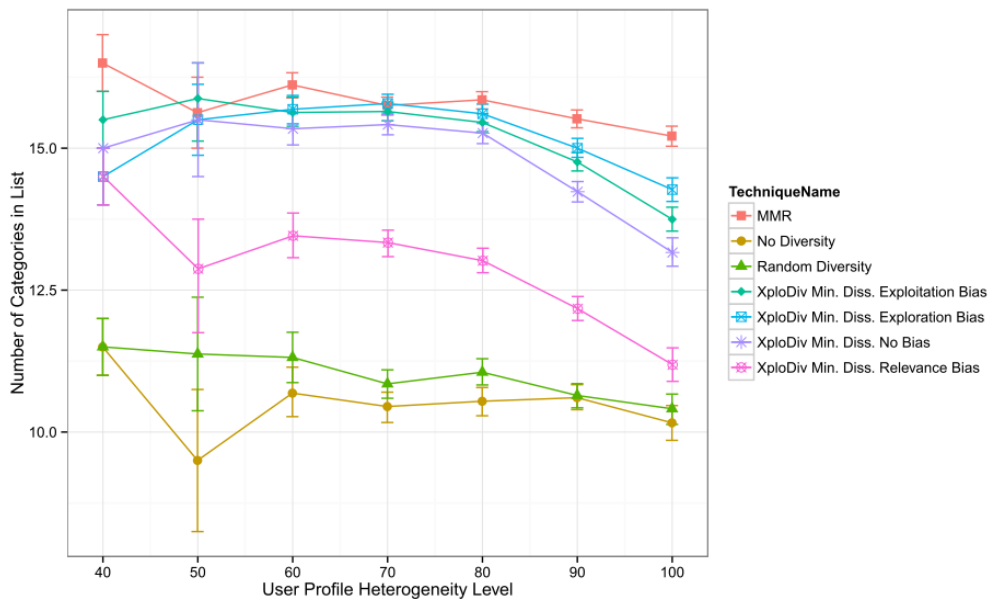


Figure 32. Number of Categories in List Metric — [Graph B] Minimum Dissimilarity Graph

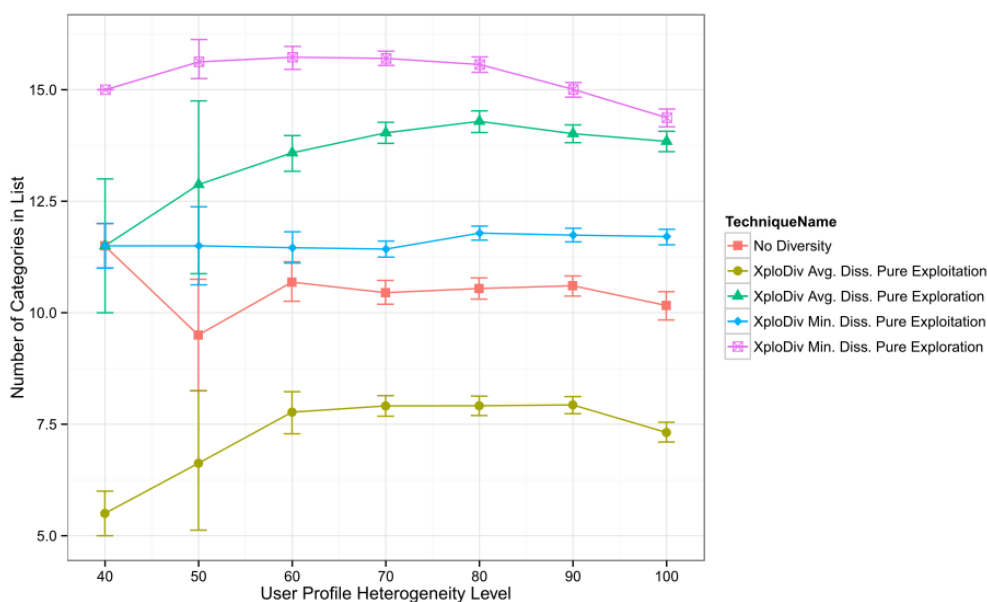


Figure 33. Number of Categories in List Metric — [Graph C] Tuning Graph

From *NumCategories* results, we observe that:

- Observing both *NumCategories* [Graph A] and *NumCategories* [Graph B], we can see that *MMR* generates result lists by giving priority to items with more associated categories.
- From *NumCategories* [Graph B], we can observe that when using the diversity dimension Minimum Dissimilarity, *XPLODIV* configurations produce results closer to *MMR*.
- We observe that No Diversity and Random Diversity have the lowest number of associated genres per item. These techniques are also very good, if not the best, at exploiting the user profile.
- From *NumCategories* [Graph C], we observe that Pure Exploration techniques have the highest number of associated genres in contrast to Pure Exploitation.
- From *NumCategories* [Graph C], we observe that Pure Exploitation using Average Dissimilarity, has the lowest number of associated categories/genres in the result list compared to all test cases (*i.e.*, even compared to results from *NumCategories* [Graph A] and *NumCategories* [Graph B]).
- From *NumCategories* [Graph C], we observe that both Pure Exploration and Pure Exploitation, when using Minimum Dissimilarity, have a higher number of associated genres in the list than No Diversity.

As a general observation, it seems that techniques that are better at exploiting the user profile have a lower number of associated categories, and that exploration oriented techniques have a high number of associated categories. Also, when using the diversity dimension Minimum Dissimilarity, items in the result set will be selected giving priority to those that have more associated genres. By giving precedence to items with more associated genres we can increment diversity when evaluated using *Gini*, as was found when analyzing the diversity perspective, as *Gini* evaluates for balance and variety. Nonetheless, when diversity was evaluated with *PILD* it was found that more associated categories lead to items within the result list to be more similar to each other.

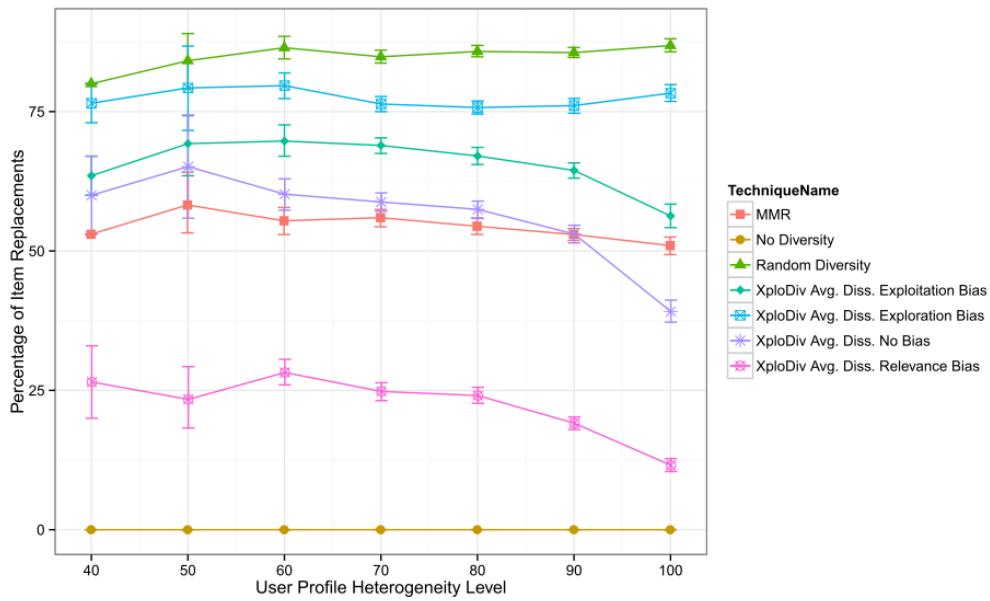


Figure 34. Percentage of Item Replacements Metric — [Graph A] Average Dissimilarity Graph

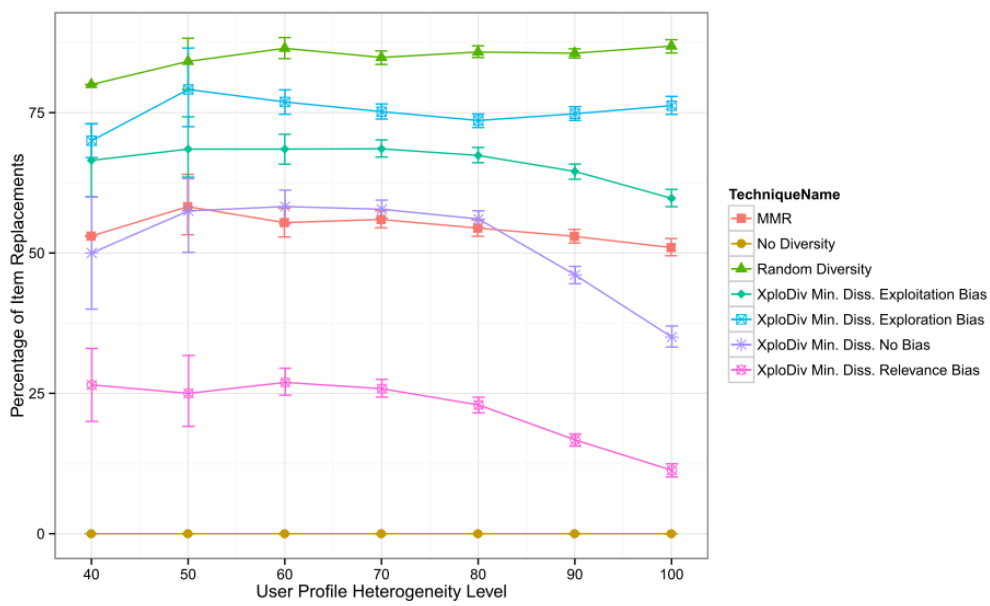


Figure 35. Percentage of Item Replacements Metric — [Graph B] Minimum Dissimilarity Graph

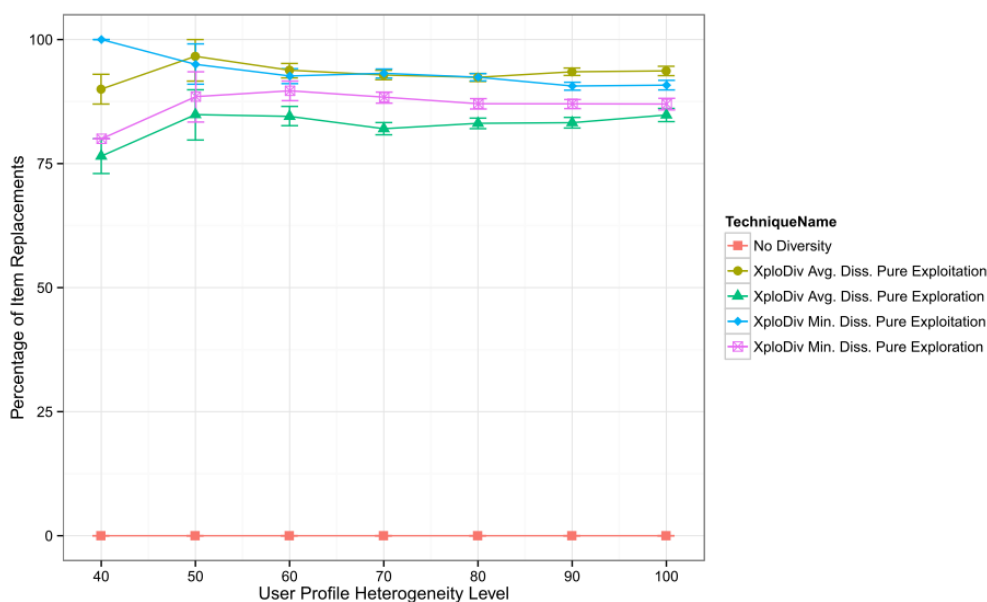


Figure 36. Percentage of Item Replacements Metric — [Graph C] Tuning Graph

From *ItemsReplaced* results, we observe that:

- Relevance Bias has the lowest number of item replacements.
- *MMR* and No Bias tend to have a similar number of item replacements. It is interesting to note that for both these approaches, the parameter that controls the diversity vs. relevance trade-off is set to 0.5, and that the number of item replacements ranges approximately between 40%-60%.
- From *ItemsReplaced* [Graph A] and *ItemsReplaced* [Graph B], we can observe that Random Diversity has the highest number of item replacements. It is followed by Exploration Bias.
- In general, *XPLODIV* approaches tend to have a higher number of item replacements compared to *MMR*.
- The diversity dimension used does not seem to have an impact on the tendencies of *ItemsReplaced*, as can be observed by comparing *ItemsReplaced* [Graph A] and *ItemsReplaced* [Graph B].
- From *ItemsReplaced* [Graph C], we observe that Pure Exploitation results have a higher number of item replacements than Pure Exploration results.

From the *ItemsReplaced* metric, we could infer that relevance oriented approaches have the lowest number of item replacements. In general, we highlight that *XPLODIV* approaches have a high number of item replacements, and therefore, a noticeable change between using or not our approach would definitely be reflected in results. After analyzing previous perspectives, it follows that this change would depend on the configuration of the approach and the parameters.

In this section, we have analyzed results for the Statistical Perspective. In the following section, we will offer a comparison of results among the analyzed perspectives by observing an integrated view. This integrated view, will better help to understand the associated trade-offs between the aspects of relevance, diversity, exploitation and exploration.

Integrated view of Evaluation results for all Perspectives

In this section, we wish to compare results from the most relevant metrics among the analyzed perspectives. Up to this point, we have individually analyzed all perspectives and shown that *XPLODIV* cannot only be tuned, but it also can be configured to produce results comparable or better than state-of-the-art approaches. In this manner, we have provided evidence to prove both *Hypothesis I* and *II*. By offering a broader view of results, in this section, we wish to offer further insight on the tunability and trade-offs between the main result list aspects, *i.e.*, relevance, diversity, exploitation and exploration.

In *Table 14*, we present results from test cases for each of the core evaluation perspectives, without considering the user profile heterogeneity. The presented results are obtained by averaging results produced for the evaluation metric by the test case over all users. We ran the *XPLODIV* test cases for both diversity dimensions —Average Dissimilarity (*Avg. Diss.*) and Minimum Dissimilarity (*Min. Diss.*)—. For each core perspective, we selected a representative metric as follows:

- (i) *Relevance Perspective: normalized Discounted Cumulative Gain (nDCG).*
- (ii) *Diversity Perspective: Pairwise Intra-List Dissimilarity metric (PILD).*
- (iii) *Exploitation Perspective: User Profile Exploitation metric (UPE).*
- (iv) *Exploration Perspective: Dissimilarity Threshold Percentage metric (DTP).*

We choose these metrics after observing results obtained from the individual analysis of the different perspectives, and considering the characteristics of the proposed metrics explained in *section 5.3.2*.

	Relevance Perspective (nDCG)	Diversity Perspective (PILD)	Exploitation Perspective (UPE)	Exploration Perspective (DTP)
<i>No Diversity</i>	1.00000	0.21232	0.58880	0.27250
<i>Random Diversity</i>	0.89090	0.23007	0.61400	0.24190
<i>MMR</i>	0.96140	0.18219	0.60690	0.49813
<i>XPLODIV Avg. Diss. Pure Exploration</i>	0.88990	0.15175	0.36650	0.69280
<i>XPLODIV Avg. Diss. Pure Exploitation</i>	0.84910	0.45730	0.65880	0.00049
<i>XPLODIV Avg. Diss. Exploration Bias</i>	0.90400	0.18324	0.36760	0.63030
<i>XPLODIV Avg. Diss. Exploitation Bias</i>	0.93050	0.25904	0.50520	0.33600
<i>XPLODIV Avg. Diss. No Bias</i>	0.95480	0.18826	0.46210	0.50270
<i>XPLODIV Avg. Diss. Relevance Bias</i>	0.99440	0.19320	0.55990	0.37060
<i>XPLODIV Min. Diss. Pure Exploration</i>	0.86760	0.22246	0.40460	0.58220
<i>XPLODIV Min. Diss. Pure Exploitation</i>	0.85440	0.50430	0.70260	0.00650
<i>XPLODIV Min. Diss. Exploration Bias</i>	0.90670	0.22860	0.42500	0.54190
<i>XPLODIV Min. Diss. Exploitation Bias</i>	0.92900	0.27014	0.54070	0.37560
<i>XPLODIV Min. Diss. No Bias</i>	0.95790	0.21680	0.52420	0.45680
<i>XPLODIV Min. Diss. Relevance Bias</i>	0.99430	0.19682	0.58440	0.36310

Table 14. Average of Results for Test Cases

In *Table 14*, we specifically compare *XPLODIV* test cases from the different diversity dimensions among themselves. We highlight with a green color, the test case result that achieves the higher value from that particular configuration. In this fashion, we made

comparisons with the following structure: *Avg. Diss.* Test Case Name vs. *Min. Diss.* Test Case Name. With these results we corroborate findings from individually analyzing perspectives, such as: *Min. Diss.* is a better diversity dimension for achieving higher diversity and exploitation, relevance is not affected by the chosen diversity dimension and *Avg. Diss.* is a better diversity dimension when aiming for high exploration. We specify that *Avg. Diss.* is better for exploration for the reason that, when results are lower compared to *Min. Diss.* is precisely in exploitation oriented approaches, which is a desirable characteristic. Because *Min. Diss.* improves over more aspects, as well as the fact that results provided by *Min. Diss.* for exploration are also successful; we choose to focus on comparing results from selected baselines and *XPLODIV*, using as diversity dimension *Min. Diss.*, in Figure 37.

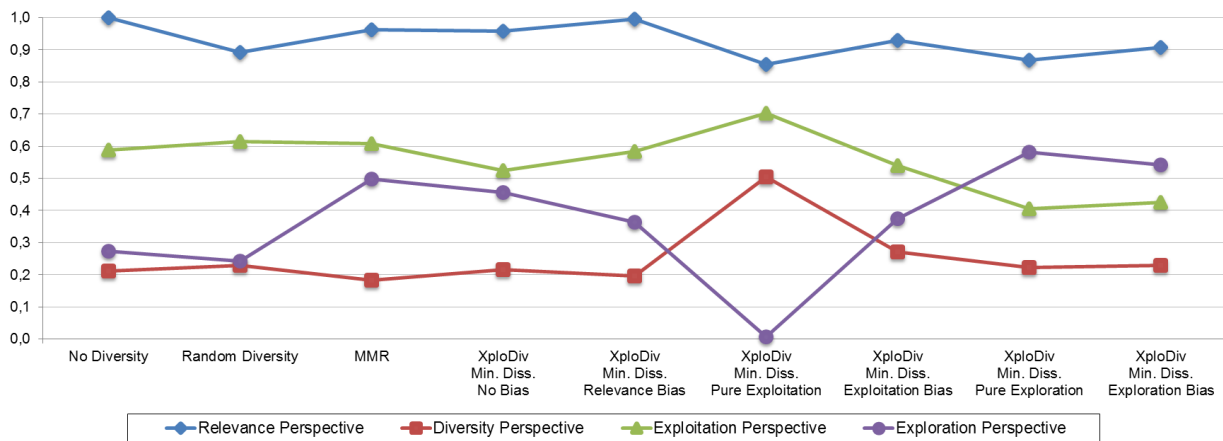


Figure 37. Integrated view of Evaluation Perspectives

In addition, in Figure 38, we present a Win-Loss diagram with respect to No Diversity, of *MMR* and *XPLODIV* test cases —using *Min. Diss.* as diversity dimension—. With this diagram, we can observe the gain or loss of each aspect with respect to No Diversity results.

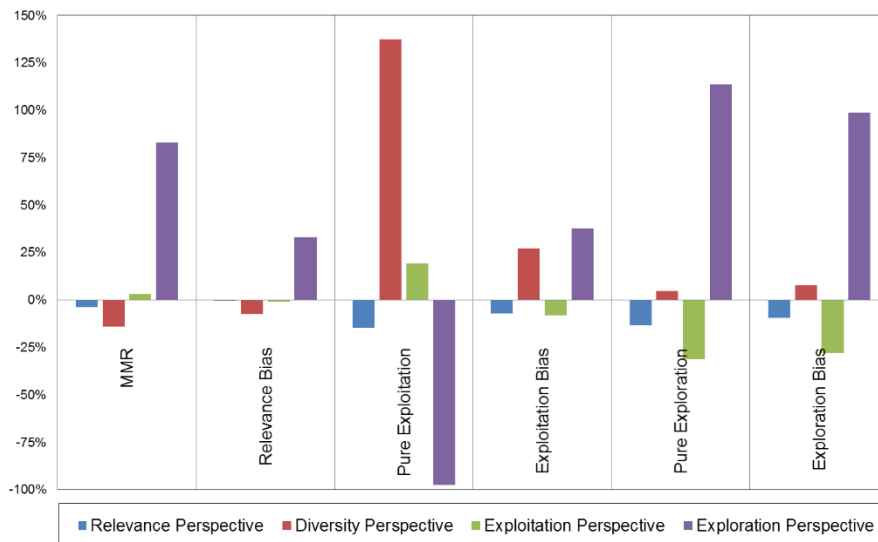


Figure 38. Win-Loss comparison of Evaluation Perspectives

From Figure 38, we can observe:

- All *XPLODIV* approaches present a diversity gain, except for the Relevance Bias. This is expected due to the trade-off between relevance vs. diversity.

- *MMR* presents both diversity and relevance loss with respect to No Diversity. However, this approach does present a notable exploration gain.
- We can observe the trade-off between relevance vs. diversity in the sense that all approaches have a relevance loss when aiming to improve diversity. However, as has been mentioned, the only approaches that do not present a relevance loss for a diversity gain are Relevance Bias and *MMR*.
- Pure Exploitation and Pure Exploration, present the largest relevance loss: 14.56% and 13.24% respectively. This is expected as for both approaches relevance was ignored to give complete precedence to diversity.
- The Relevance Bias approach produces the lowest relevance loss of 0.57%.
- The largest diversity gain was obtained by Pure Exploitation: diversity gain of 137.52%. This configuration also had the largest exploration loss: exploration loss of 97.61%. This loss can be explained by the exploitation vs. exploration trade-off.
- Our approach can be tuned towards higher exploration, as can be observed by the large gain in the Exploration Perspective obtained by the Pure Exploration and Exploration Bias approaches. Even though *MMR* presents a large exploration gain, *XPLODIV* exploration oriented approaches offer larger exploration gain.
- The largest exploration gain was obtained by Pure Exploration: exploration gain of 113.65%.
- The largest exploitation gain was obtained by Pure Exploitation: exploitation gain of 19.33%. All other configurations, except for *MMR*, obtained a small exploitation loss. This could be further explained by analyzing a possible interdependent relation between the relevance and exploitation aspects.
- The Exploitation Bias approach, has both a small exploitation and relevance loss, for a noteworthy gain in exploration and diversity.
- It could be observed that when there was a large exploration gain, the obtained diversity gain was not considerably high, even in some occasions having a diversity loss when an exploration gain was achieved, such as in: *MMR* and Relevance Bias. In the Pure Exploitation approach we can observe a large diversity gain but also a large exploration loss. It would be interesting to analyze the similarity among explorative items in the candidate items list to see if adding explorative items is actually affecting the diversity outcome.

Similar results can be viewed from *Figure 37*, in addition we highlight:

- If we observe the Exploitation Perspective and Relevance Perspective tendencies, when exploitation values go up relevance tends to go down.
- Pure Exploitation presents: the lowest relevance value, the largest diversity value, the highest exploitation value and the smallest exploration value.
- The lowest exploitation values are produced by exploration oriented approaches of *XPLODIV*.
- The lowest exploration values are not necessarily produced by having a high exploitation value. The lowest values of exploration are obtained by No Diversity, Random Diversity and Pure Exploitation. This leads to the idea that a post-filtering selection method, of a final recommendation list from candidate items, is still needed if a traditional *RecSys* wishes to provide users novel items, regardless of wanting to offer or not diversification.

In this section, we have analyzed an integrated view of results obtained from the different evaluation perspectives. With these results, we offer further evidence to prove *Hypothesis I* and *Hypothesis II*. In addition, we observed hints of possible additional relations between the aspects, such as the connection between exploitation and relevance, and a possible trade-off between diversity and exploration. To further analyze our intuitions, further studies could be carried out to analyze the diversity of explorative products within the candidate item set to see if they similar to each other or not. Also, we could analyze the connection between relevance and exploitation in different environments by using other *RecSys* algorithms to generate candidate items and even other datasets.

In the following section, we will present a summary of the presented Experimental Validation.

5.4 Summary

In this section, we have presented Experimental Validation to prove that *XPLODIV* can: (i) be tuned using the control parameters towards results with greater relevance, diverse exploitation or diverse exploration, and (ii) generate results comparable, and in given test cases better, than baselines and state-of-the-art approaches. In order to argument towards these hypothesis, we carried out both qualitative and quantitative tests. Within these tests, we also explored the influence that the heterogeneity of the user profile had over results. Because *XPLODIV* can be configured in different ways, it was very important to identify the dimensions to be used in each test case, particularly we gave special attention to the analysis the impact over results of using a different diversity dimension.

In qualitative tests, we selected three users with different levels of heterogeneity, and analyzed results generated by the different test configurations for these users. We carried out experiments on two scenarios, pure exploitation and pure exploration, and in each, we proposed test configurations for *XPLODIV* considering the possible combinations of exploitation dimensions and exploration dimensions. With qualitative tests, we wanted to observe the tunability of *XPLODIV*, and therefore we did not analyze baselines or state-of-the-art techniques. Through qualitative observations of test outputs, we highlighted that our approach could be tuned to be purely explorative or purely exploitative on the different combinations of available dimensions. From results, we concluded that for the MovieLens dataset, the dimensions *ImportanceOfAssociatedPreference* and *UserProfileNovelty* excel as exploitation dimension and exploration dimension respectively (reference *Table 9*, to view implementation details for used dimensions). For this reason, we used these dimensions to configure *XPLODIV* for quantitative tests. As for the diversity dimension, in quantitative tests, we continued to analyze both possible instantiations: *AverageDissimilarity* and *MinimumDissimilarity*.

In quantitative tests, we evaluated the described configuration of our approach mainly from four different perspectives, showing that it not only provides comparable results to state-of-the-art techniques, but in addition, it can be tuned towards more exploitative diversity or more explorative diversity. First and foremost, in order to define a methodology for quantitative tests, we defined a Diversity-Aware Evaluation Framework in which we structure perspectives and propose metrics to evaluate each perspective. The analyzed perspectives measure the relevance, diversity, exploitation and exploration aspects of obtained results. In general, we observed results from each perspective and then offered an integrated view of results.

To conclude, we offer answers to the questions identified in the introduction of the Experimental Validation in *Table 15*.

Questions	Answers
<i>Can the trade-offs (i.e., relevance vs. diversity and exploitation vs. exploration) be observed?</i>	We can most clearly observe these trade-offs when analyzing the integrated view of perspectives in quantitative tests, specifically in the win-loss graph presented in <i>Figure 38</i> . Nonetheless, when analyzing the individual perspectives we constantly presented evidence to describe the nature of these trade-offs.
<i>Can the parameters α and β control the characteristics of generated results in terms of relevance, diversity, exploitation and exploration?</i>	In both quantitative and qualitative tests, given different configurations <i>XPLODIV</i> , we showed that our approach could be tuned using the parameters α and β .
<i>In which scenarios does <i>XPLODIV</i> outperform baseline techniques?</i>	We identify that <i>XPLODIV</i> outperforms baselines when offering explorative results. We found that all <i>XPLODIV</i> configurations, except for Pure Exploitation, outperformed Random Diversity and No Diversity in terms of exploration. Also, <i>XPLODIV</i> configurations excelled in diversity except for Relevance Bias.
<i>In which scenarios does <i>XPLODIV</i> perform worse than baselines techniques?</i>	We found that <i>XPLODIV</i> presented unexpected results in the Exploitation Bias configuration. Though results were unexpected, we have yet to determine if they are necessarily worse than baselines. In detail, all techniques present a relevance loss in comparison to No Diversity, which is expected due to the trade-off between diversity vs. relevance. However, in terms of exploitation, the Exploitation Bias configuration had lower exploitation value than No Diversity. However, the Exploitation Bias configuration had a larger diversity and exploration value than No Diversity in trade-off to the exploitation loss.
<i>Do different configurations of <i>XPLODIV</i> affect outcomes?</i>	Yes, different configurations of <i>XPLODIV</i> have an impact over generated results. Specifically in qualitative tests, by analyzing the different combinations of the possible dimensions (i.e., exploration, exploitation and diversity dimensions), we found that different dimensions generated very different results. Also, in quantitative tests, we found that certain diversity dimensions performed better for different objectives.
<i>What additional trade-offs can be observed?</i>	As for additional trade-offs, when analyzing quantitative tests, we found that there is a possible positive correlation between relevance and exploitation, and a possible trade-off between exploration and diversity. Further studies are needed to corroborate these intuitions.

Table 15. Answers to Experimental Validation Questions

In this section, we have provided robust qualitative and quantitative validation to show that *XPLODIV* produces comparable results to baselines and state-of-the-art techniques; while in addition, providing control over the amount of desired relevance, diversity, exploitation and

exploration that can be found in generated results. In the following section, we conclude our work and present perspectives for further research.

Chapter VI

CONCLUSIONS AND FUTURE WORK

In this work we have addressed the problem of diversification in the field of Recommendation Systems. We found that, even though diversity is a desirable feature, it is a challenge for traditional Recommendation Systems to offer diverse recommendations, we highlight the following reasons: (a) the heuristics that lay foundation to recommendation techniques are based on similarity measures that limit the diversity of items considered for recommendation, (b) evaluation metrics that assess the individual quality of items in Recommendation Systems penalize diversity and novelty, and (c) recommendation list evaluation is performed as an aggregate of the individual scores of items, disregarding the real value of items in the context of the list.

From our literature review, we found that current diversification techniques for Recommender Systems are mostly inspired from works in Information Retrieval. For this reason, current diversification techniques, in general, disregard the concept of exploration, some even penalizing novel products in the selection process for the final recommendation list. We highlight that exploration of novel products should be explicitly considered, as this notion is essential to discovery, which in turn, is the most important feature of a Recommender System. For this reason, we define the following research goal for our work:

Design a diversification technique for *RecSys* that can balance the trade-off between quality (in terms of relevance) and diversity, considering the trade-off between *exploitation* of the user profile and *exploration* of novel products.

As a solution, in this work, we propose the Exploitation-Exploration Diversification technique named *XPLODIV*. Specifically, we highlight the following main contributions:

- (i) ***Exploitation-Exploration diversification approach (XPLODIV)***: we present a novel diversification technique called *XPLODIV*, which considers not only the trade-off between relevance *vs.* diversity, but also the trade-off between exploitation *vs.* exploration. This technique is composed of four core dimensions, which are: Relevance Dimension, Diversity Dimension, Exploitation Dimension and Exploration Dimensions. For each dimension, we provide a detailed description of what the dimension measures and offer alternative methods to achieve the dimension's goal. Our experimental evaluation showed that the proposed approach generates comparable results to baseline and state-of-the-art techniques. Moreover, through control parameters, our approach can be tuned towards more explorative or exploitative recommendations. We emphasize that *XPLODIV*, explicitly considers the aspect of exploration, which is disregarded in current works. The exploration aspect is crucial to *RecSys*, given its direct influence on factors related to novelty, serendipity and discovery.
- (ii) ***Diversity-Aware Evaluation Framework for Recommender Systems***: we introduce a Diversity-Aware Evaluation Framework, which identifies and organizes metrics that should be taken into account when evaluating Recommendation Systems within the context of diversity. Our framework allows to evaluate results from four core perspectives, which are: Relevance Perspective, Diversity Perspective, Exploitation Perspective and Exploration

Perspective. For each perspective, we associate a number of metrics from current work and novel metrics proposed in this work. In order to obtain a complete view of the quality of results, perspectives should be first analyzed individually, and then in comparison to each other.

- (iii) **Analytical comparison of Related Work:** we carry out a robust analytical comparison of related work, which to our knowledge, has not been carried out before. Findings from our literature review improve knowledge on the field of diversification for Recommender Systems, by emphasizing the advantages and disadvantages of current approaches, and serve as valuable reference for future work.

In conclusion, it was found that our proposed approach —*XPLODIV*— satisfies the research goal, by defining a diversification technique that can balance not only the trade-off between relevance and diversity, but also the trade-off between exploitation of the user profile and exploration of novel products. Furthermore, our technique is an improvement over current work, because in addition to providing comparable results, it can be tuned towards more diverse explorative results or more diverse exploitative results. In this fashion, *XPLODIV* can be adjusted to the Recommender System requirements of relevance, diversity, exploitation and exploration.

Given the short amount of time, a lot was accomplished, however future work is still needed to further explore the potential of the proposed diversification technique. Some open questions have already been mentioned across this document. Nonetheless, subsequently, we discuss the most relevant directions for future research which could not be covered by the scope of this project:

Adaptation of diversity vs. relevance and exploitation vs. exploration trade-off parameters

The required level of diversity, exploitation and exploration could depend on many factors (e.g., heterogeneity of the user profile, context, product domain, among others). It is important to define these external factors and study their influence on the parameters that define the diversity vs. relevance and exploitation vs. exploration trade-off parameters. This could lead to answer questions such as: is the user profile heterogeneity level correlated to the user's diversity, exploitation or exploration needs? In addition, the adaptation of these parameters over time should also be studied in relation to the changing needs of the user. Moreover, parameters should adapt in accordance to the user's implicit or explicit feedback over time. That is to say, the way the user interacts with the received recommendation list could offer information on the users diversity, exploration or exploitation needs that should be used to adapt the parameters. For example, the user could explicitly indicate to the *RecSys* that he/she wants to receive more diverse results, more explorative results or more exploitative results given the options in the user interface.

The use of XPLODIV as a technique to enhance traditional Recommendation Algorithms

The proposed diversification technique *XPLODIV*, could be applied not only as a post-filtering diversification technique but also as a technique used to enhance current *RecSys* algorithms. In the literature review, we analyzed works that used diversification techniques in order to select a diverse set of neighbors in a collaborative filtering algorithm. Similar studies could be carried out with *XPLODIV* to answer questions such as: what is the exploitative value of a particular neighbor? what is the exploration value of a particular neighbor? and can diversifying the user

neighbors with *XPLODIV* give control over the amount of exploitative or explorative items that can be found in the candidate items?. In addition, combined approaches can be studied, where the candidate items for the post-filtering *XPLODIV* could come from an *XPLODIV* enhanced *RecSys* algorithm.

The use of XPLODIV as an aggregation technique in hybrid Recommendation Systems

Different *RecSys* algorithms produce results with different characteristics. One way to have a hybrid *RecSys* is just to aggregate the results from two or more *RecSys* algorithms. We would like to study the impact on *XPLODIV* results if the candidate items were obtained from different *RecSys* algorithms.

The impact of XPLODIV on real user satisfaction

Online evaluations of *XPLODIV* would give a greater insight on the impact of considering diversity, exploitation and exploration on user satisfaction. The set-up of a study of this magnitude was outside of the scope of the current project. However, this would lead to answer questions such as: what is the user's reaction to different diversification techniques? and, can users perceive that results are more exploitative, explorative or diverse?

How does the recommendation technique used affect the results of XPLODIV

For this work, *XPLODIV* was evaluated with candidate items generated from one *RecSys* algorithm. However, in the future, it is important to study the diversity and novelty level of results offered by different *RecSys* algorithms, and how these affect the results of the post-filtering technique *XPLODIV*. It would also be interesting to view how the results from diversification enhanced recommenders, such as the reviewed in *section 3.1.2.7*, impact the results of *XPLODIV*.

The Diversification Problem in Non-Traditional Recommendation Systems

In this work, we have studied the diversification problem in traditional *RecSys*. However, non-traditional *RecSys*, could have different diversification needs. This future direction would examine the question of: Can *XPLODIV* be adapted to be used on non-traditional Recommendation Systems?

In general, non-traditional approaches bend the at least one of the "rules" indicated for traditional *RecSys*, such as: offer one recommendation, to one user, solely based on rating information found in the User-Item matrix, where the relation between an item and a user is represented by a single-valued numerical rating. Examples of non-traditional *RecSys* are: group recommenders (offer one suggestion to be shared among more than one user) or multi-criteria recommenders (consider more than one user rating to describe items).

As an example, a possible research focus could be on the impact of diversification in stream-based *RecSys*, where items are suggested in sequence, such as in a personalized music radio station. In the case of the music radio, we might want diversity in the long term, but we would like results in the short term not to be so different from each other and therefore maintain the "flow" between the reproduced songs.

Measuring dissimilarity between categories

This work highlighted that categories could have inter-dependent relationships. However, it is not easy to measure the level of dissimilarity between two categories (*e.g.*, how similar are the

genres Drama and Comedy?). Future work would extend on these measurements and view the impact of using category dissimilarity information on the result of diversification techniques.

Adapt XPLODIV to enhance inter-list diversification

It is not a desirable feature for the Recommendation System to show the user the same list of diversified results over and over again. Diversity must also be offered over time. It is important to adjust *XPLODIV* to augment inter-list diversification and not only intra-list diversification.

REFERENCES

- [Adom05] Adomavicius, G., Tuzhilin, A. Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions. In: IEEE Transactions on Knowledge and Data Engineering, vol. 17, no. 6, June 2005. IEEE Educational Activities Department (Piscataway, NJ, USA, 2005), pp. 734-749.
- [Adom08] Adomavicius, G., Huang, Z., Tuzhilin, A. Personalization and Recommender Systems. In: Zhi-Long, Z., Raghavan, S., Gray, P. (eds.): Tutorials in Operations Research 2008, State-of-Art Decision Making Tools in the Information-Intensive Age, INFORMS 2008, October 12-15, 2008. Institute for Operations Research and the Management Sciences (INFORMS), pp. 55-105.
- [Adom09] Adomavicius, G., Kwon, Y. Toward more diverse recommendations: Item re-ranking methods for recommender systems. (eds.): Proceedings of the 19th Workshop on Information Technology and Systems (WITS 2009) (Phoenix, Arizona, December 14-15, 2009).
- [Adom12] Adomavicius, G., & Kwon, Y. Improving Aggregate Recommendation Diversity Using Ranking-Based Techniques. IEEE Transactions on Knowledge and Data Engineering, 24(5) (2012), 896-911.
- [Agra09] Agrawal, R., Gollapudi, S., Halverson, A., Ieong, S. Diversifying Search Results. In Proceedings of the Second ACM International Conference on Web Search and Data Mining (2009) (pp. 5-14). New York, NY, USA: ACM.
- [Akin12] Akinyemi, J. Similarity and Diversity in Information Retrieval (2012, April 30). University of Waterloo
- [Ande06b] Anderson, C. The Long Tail: Why the Future of Business is Selling Less of More. Hyperion (July 11, 2006), 256 pages.
- [Apac14] What is Apache Mahout?: <http://mahout.apache.org/> (Last Access: November 2014).
- [Baez99] Baeza-Yates, R., Ribeiro-Neto, B. Modern Information Retrieval, 1st edition. Addison Wesley (May 15 1999), 544 pages.
- [Bill00] Billsus, D., Pazzani, M. User Modeling for Adaptive News Access. In: User Modeling and User-Adapted Interaction, vol. 10, nos. 2-3, June 2000. Kluwer Academic

Publishers (Netherlands), pp. 147-180, 2000.

- [Burk07] Burke, R. Hybrid Web Recommender Systems. In: Brusilovsky, P., Kobsa, A., Nejdl, W. (eds.): *The Adaptive Web*. Springer, Berlin, Heidelberg (2007), pp. 377-408.
- [Carb98] Carbonell, J. Goldstein, J. The Use of MMR, Diversity-Based Reranking for Reordering Documents and Producing Summaries. In: *SIGIR'98*, Melbourne, Australia. ACM.
- [Cart11] Carterette, B. An analysis of NP-completeness in novelty and diversity ranking. *Information Retrieval*, (2011),14(1), 89-106.
- [Ceri13] Ceri, S., Bozzon, A., Brambilla, M., Valle, E. D., Fraternali, P., Quarteroni, S. Recommendation and Diversification for the Web. (eds.): *Proceedings of Web Information Retrieval* (pp. 111-120) (2013). Springer Berlin Heidelberg.
- [Chen09] Chen, S., Ma, B., Zhang, K. On the similarity metric and the distance metric. *Theoretical Computer Science*, (2009),410(24-25), 2365-2376.
- [Chen13] Chen, J., Liu, Y., Hu, J., He, W., Li, D. A Novel Framework for Improving Recommender Diversity. In: Cao, L., Motoda, H., Srivastava, J., Lim, E-P., King I., Yu, P, Nejdl, W., Xu, G., Li, G. (eds.): *Behavior and Social Computing, Lecture Notes in Computer Science*, vol., 8178 (2013), pp. 129-138.
- [Clar08] Clarke, C. L. A., Kolla, M., Cormack, G. V., Vechtomova, O., Ashkan, A., Bütcher, S., MacKinnon, I. Novelty and Diversity in Information Retrieval Evaluation. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (2008)(pp. 659-666). New York, NY, USA: ACM.
- [Clar09] Clarke, C. L., Kolla, M., Vechtomova, O. An effectiveness measure for ambiguous and underspecified queries (2009). Springer.
- [Dros10] Drosou, M., Pitoura, E. Search Result Diversification. In *Proceedings of SIGMOD Rec. 2010*, 39(1), 41-47.
- [GilC13] Gil-Costa, V., Santos, R. L. T., Macdonald, C., Ounis, I. Modelling Efficient Novelty-based Search Result Diversification in Metric Spaces. *J. of Discrete Algorithms* (2013), 18, 75-88.
- [Goll09] Gollapudi, S., Sharma, A. An Axiomatic Approach for Result Diversification. In *Proceedings of the 18th International Conference on World Wide Web* (2009).(pp. 381-390). New York, NY, USA: ACM.

- [Gord92] Gordon, M. D., Lenk, P.. When is the probability ranking principle suboptimal? *Journal of the American Society for Information Science* (1992), 43(1), 1-14.
- [Grou14] MovieLens Datasets: <http://grouplens.org/datasets/movielens/> (Last Access: November 2014).
- [Herl04] Herlocker, J.L., Konstan, J.A., Terveen, L.G., Riedl, J.T.. Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems*, vol., 22, issue, 1, (January 2004), pp. 5-53.
- [Hurl11] Hurley, N., Zhang, M. Novelty and Diversity in Top-N Recommendation - Analysis and Evaluation. In: *ACM Transactions on Internet Technology (TOIT)*, vol., 10 issue, 4, no., 14, March 2011. ACM New York, NY, USA.
- [Jung94] Junge, K. Diversity of ideas about diversity measurement. *Scandinavian Journal of Psychology*(1994), 35(1), 16-26.
- [Lath10] Lathia, N., Hailes, S., Capra, L., & Amatriain, X. (2010). Temporal Diversity in Recommender Systems. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 210-217). New York, NY, USA: ACM.
- [Lemi08] Lemire, D., Downes, S., Paquet, S. Diversity in open social networks. Published online. 2008
- [Macl08] Maclaurin, J., & Sterelny, K. *What Is Biodiversity?* University of Chicago Press(2008).
- [Mank04] Mankiw, N. G. *Principles of Economics*, 3rd Edition. 2004. Cengage Learning.
- [Maro60] Maron, M. E., & Kuhns, J. L. On Relevance, Probabilistic Indexing and Information Retrieval. *Journal of the ACM* (1960), 7(3), 216-244.
- [Mcdo03] McDonald, D. G., Dimmick, J. The Conceptualization and Measurement of Diversity(2003). *Communication Research*, 30(1), 60-79.
- [Mcne06] Mcnee, S.M., Riedl, J., Konstan, J.A.. Being accurate is not enough: how accuracy metrics have hurt recommender systems. In: Olson, G., Jeffries, R. (eds.): *Proceedings of the 2006 Conference on Human Factors in Computing Systems (CHI 2006)* (Montréal, Québec, Canada, April 22-27, 2006). ACM, New York (2006), pp. 1097-1101.
- [McSh02] McSherry, D. Diversity-Conscious Retrieval. In S. Craw & A. Preece (Eds.), *Advances in Case-Based Reasoning*(2002) (pp. 219-233). Springer Berlin Heidelberg.

- [Morv98] Morville, P. Information Architecture on the World Wide Web, 1 ed., February 1998. O'Reilly Media (1998), 202 pages.
- [Nehr02] Nehring, K., & Puppe, C. (2002). A Theory of Diversity. *Econometrica*, 70(3), 1155-1198.
- [Nem78] Nemhauser, G. L., Wolsey, L. A., Fisher, M. L. An analysis of approximations for maximizing submodular set functions-I. *Mathematical Programming*, (1978), 14(1), 265-294.
- [Nemh78] Nemhauser, G. L., Wolsey, L. A., Fisher, M. L. An analysis of approximations for maximizing submodular set functions-I. *Mathematical Programming* (1978), 14(1), 265-294.
- [Owen12] Sean Owen, Robin Anil, Ted Dunning, and Ellen Friedman. 2011. Mahout in Action. Manning Publications Co., Greenwich, CT, USA.
- [Pati82] Patil, G. P., & Taillie, C. Diversity as a Concept and its Measurement. *Journal of the American Statistical Association* (1982), 77(379), 548-561.
- [Radl09] Radlinski, F., Bennett, P. N., Carterette, B., Joachims, T. Redundancy, Diversity and Interdependent Document Relevance. *SIGIR Forum*, (2009), 43(2), 46-52.
- [Robe77] Robertson, S. The probability ranking principle in IR. *Journal of Documentation*, 33(4),(1977), 294-304.
- [Said12] Said, A., Jain, B., Kille, B., Albayrak, S. Increasing Diversity Through Furthest Neighbor-Based Recommendation. Presented at the Proceedings of the WSDM'12 Workshop on Diversity in Document Retrieval (2012)(DDR'12).
- [Sant10] Santos, R. L. T., Macdonald, C., & Ounis, I. Exploiting Query Reformulations for Web Search Result Diversification. In *Proceedings of the 19th International Conference on World Wide Web (2010)*(pp. 881-890). New York, NY, USA: ACM.
- [Sant10a] Santos, R. L. T., Macdonald, C., Ounis, I. Selectively Diversifying Web Search Results. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management* (pp. 1179-1188) (2010).. New York, NY, USA: ACM.
- [Shi12] Shi, Y., Zhao, X., Wang, J., Larson, M., & Hanjalic, A. Adaptive Diversification of Recommendation Results via Latent Factor Portfolio. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval (2012)*(pp. 175-184). New York, NY, USA: ACM.

- [Smyt01] Smyth, B., McClave, P. Similarity vs. Diversity. In D. W. Aha & I. Watson (Eds.), *Case-Based Reasoning Research and Development* (2001)(pp. 347-361). Springer Berlin Heidelberg.
- [Stir07] Stirling, A. A General Framework for Analysing Diversity in Science, Technology and Society. SPRU Working Paper Series No. 156, (2007). SPRU - Science and Technology Policy Research, University of Sussex.
- [Stir98] Stirling, A. On the economics and analysis of diversity (1998, January 1)[Monograph].
- [Tint13] Tintarev, N., Dennis, M., Masthoff, J. Adapting Recommendation Diversity to Openness to Experience: A Study of Human Behaviour. In: Carberry, S., Weibelzahl, S., Micarelli, A., Semeraro, G. (eds.): *Proceedings of the 21st International Conference on User Modeling, Adaptation, and Personalization (UMAP 2013)* (Rome, Italy, June 10-14, 2013). Springer, Berlin, Heidelberg (2013), pp. 190-202.
- [Varg11] Vargas, S., Castells, P. Rank and relevance in novelty and diversity metrics for recommender systems. In: Mobasher, B., Burke, R. (eds.): *Proceedings of the 5th ACM conference on Recommender Systems (RecSys 2011)* (Chicago, IL, USA, October 23-27,2011). ACM, New York (2011), pp. 109-116.
- [Varg12] Vargas, S. Novelty and Diversity Enhancement and Evaluation in Recommender Systems (2012, April). Universidad Autónoma de Madrid, Madrid, España.
- [Varg13] Vargas, S., Castells, P. Exploiting the Diversity of User Preferences for Recommendation. In *Proceedings of the 10th Conference on Open Research Areas in Information Retrieval* (pp. 129-136). Paris, France.
- [Varg14] Vargas, S., Baltrunas, L., Karatzoglou, A., Castells, P. Coverage, redundancy and size-awareness in genre diversity for recommender systems (2014). (pp. 209-216). ACM Press.
- [Wang09] Wang, J., & Zhu, J. Portfolio Theory of Information Retrieval. (eds.): *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 115-122) (2009). New York, NY, USA: ACM.
- [Wang09] Wang, J., & Zhu, J. (2009). Portfolio Theory of Information Retrieval. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 115-122). New York, NY, USA: ACM.
- [Weit92] Weitzman, M. L. On Diversity. *The Quarterly Journal of Economics* (1992), 107(2), 363-405.

- [Yang13] Yang, C., Ai, C. C., Li, R. Neighbor Diversification-Based Collaborative Filtering for Improving Recommendation Lists. In Proceedings of IEEE 10th International Conference on High Performance Computing and Communications (2013)(pp. 1658-1664).
- [Zhai03] Zhai, C. X., Cohen, W. W., Lafferty, J. Beyond Independent Relevance: Methods and Evaluation Metrics for Subtopic Retrieval. In Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval (2003)(pp. 10-17). New York, NY, USA: ACM.
- [Zhan08] Zhang, M., Hurley, N.. Avoiding Monotony: Improving the Diversity of Recommendation Lists. (eds.): Proceedings of the 2008 ACM Conference on Recommender Systems (RecSys 2008), (Lausanne, Switzerland, October 23 - 25, 2008). ACM, New York, NY, USA (2008), pp. 123-130.
- [Zhan09] Zhang, F. Improving recommendation lists through neighbor diversification. In IEEE International Conference on Intelligent Computing and Intelligent Systems, 2009. ICIS 2009 (Vol. 3, pp. 222-225).
- [Zhan09a] Zhang, M., & Hurley, N. Novel Item Recommendation by User Profile Partitioning. In IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technologies, 2009. WI-IAT '09 (Vol. 1, pp. 508-515).
- [Zhen12] Zheng, W., Wang, X., Fang, H., Cheng, H.. Coverage-based search result diversification. Information Retrieval, (2012), 15(5), 433-457.
- [Zieg05] Ziegler, C.N., McNee, S.M., Konstan, J.A., Lausen, G. Improving recommendation lists through topic diversification. In: Ellis, A., Hagino, T. (eds.): Proceedings of the 14th international conference on World Wide Web (WWW 2005) (Chiba, Japan, May 10-14, 2005). ACM, New York (2005), pp. 22-32.
- [Zucc12] Zuccon, G., Azzopardi, L., Zhang, D., Wang, J. Top-k Retrieval Using Facility Location Analysis. In R. Baeza-Yates, A. P. de Vries, H. Zaragoza, B. B. Cambazoglu, V. Murdock, R. Lempel, & F. Silvestri (2012). (Eds.), Advances in Information Retrieval (pp. 305-316). Springer Berlin Heidelberg.