Escuela Politécnica Superior

Departamento de Ingeniería Informática

NOVELTY AND DIVERSITY EVALUATION AND ENHANCEMENT IN RECOMMENDER SYSTEMS

PhD DISSERTATION

Saúl Vargas Sandoval

Madrid, February 2015

PhD thesis title: Novelty and Diversity Evaluation and Enhancement in Recommender Systems

Author: **Saúl Vargas Sandoval**

Affiliation: Departamento de Ingeniería Informática
Escuela Politécnica Superior
Universidad Autónoma de Madrid, Spain

Supervisor: **Pablo Castells Azpilicueta**
Universidad Autónoma de Madrid, Spain

Date: February 2015

Committee: **President:**

Secretary:

Vocal 1:

Vocal 2:

Vocal 3:

To María Pilar, Carlos and Sheila

ABSTRACT

Recommender Systems have become a pervasive technology in a wide spectrum of everyday applications, and can be said to be familiar to the general public. In situations where there is an *information overload*, such as e-commerce, streaming platforms or social networks, providing personalized recommendations has proven to be a major source of enhanced functionality, user satisfaction, and revenue improvements. The development of recommendation algorithms and technologies has typically focused on maximizing the prediction accuracy of the user's interests. However, there is an increasing awareness in the field that there are other properties that have an impact on user satisfaction and business performance. In particular, there are many cases where applying some degree of novelty or diversity may be beneficial for both the users that receive the recommendations and the business that provides them.

In this thesis we develop a principled approach to the evaluation and enhancement of novelty and diversity in Recommender Systems. We consider that the improvement of such fundamental dimensions of the usefulness of recommendations has to take into account how users explore and perceive recommendations, what are the problems that novelty and diversity solve and the causes of such problems. We propose in our first contribution a unified framework for the evaluation and enhancement of novelty and diversity in recommendations that generalizes and enhances many of the proposals previously studied in the state of the art under a common basis. Special emphasis is done in the study of the diversity within recommendations lists, for which two different contributions are presented. On the one hand, an adaptation of search result diversification metrics and techniques from Information Retrieval is explored to cope with the ambiguity of user interests and tastes. On the other hand, a domain-specific solution for assessing and optimizing the diversity of recommendations is proposed to address the need of users for varied recommendations when genre information about the recommendation domain is available. Finally, we address diversity as an overall quality from the system point of view, and we propose solutions for the problem in this perspective by turning the recommendation task around and recommending users to items.

Our proposals are tested on a common experimental design that considers three different datasets for movie and music recommendation and four well-known baseline recommendation algorithms. The results of our experiments support the validity of our contributions and allow the analysis and further insights on their behavior when applied to different settings.

RESUMEN

Los Sistemas de Recomendación se han convertido en una tecnología presente en un amplio espectro de aplicaciones de uso cotidiano, y se puede decir que son hoy en día un concepto familiar para el público en general. En situaciones donde hay una *sobrecarga de información*, como es el caso de las plataformas de comercio electrónico y *streaming* y de las redes sociales, proporcionar recomendaciones personalizadas ha demostrado ser una fuente importante de mejoras de funcionalidad, satisfacción de los usuarios y rendimiento del negocio. El desarrollo de algoritmos y tecnologías de recomendación se ha centrado tradicionalmente en maximizar el acierto en la predicción de los intereses del usuario. Sin embargo, hay una percepción general en el área de los Sistemas de Recomendación de que hay otras propiedades que tienen un impacto importante en la satisfacción del usuario y el desempeño de negocio. En particular, hay muchos casos donde aplicar un cierto grado de novedad o diversidad puede ser beneficioso tanto para los usuarios que reciben las recomendaciones como para el negocio que las provee.

En esta tesis desarrollamos un enfoque fundamentado de la evaluación y mejora de novedad y diversidad en Sistemas de Recomendación. Consideramos que la mejora de tales dimensiones fundamentales de la utilidad de las recomendaciones tiene que tener en cuenta cómo los usuarios exploran y perciben las recomendaciones, cuáles son los problemas que la novedad y la diversidad resuelven, y las causas de los mismos. En nuestra primera contribución proponemos un marco unificado para la evaluación y mejora de novedad y diversidad en recomendaciones que unifica, generaliza y refina muchas de las propuestas previamente estudiadas en trabajo previo sobre una base común. Hemos hecho asimismo un énfasis especial en el estudio de la diversidad de listas de recomendación, para la cual presentamos dos contribuciones. Por un lado, se explora una adaptación de las métricas y técnicas de diversificación de resultados de búsqueda en Recuperación de Información para lidiar con la ambigüedad de los usuarios en sus intereses y gustos. Por otro lado, se propone una solución específica al dominio para abordar la necesidad de recomendaciones variadas cuando se dispone de información de géneros en el dominio de recomendación. Por último, abordamos la diversidad como una cualidad general desde el punto de vista del sistema, y proponemos soluciones para esta perspectiva dándole la vuelta a la tarea de recomendación, recomendando usuarios a artículos, a la inversa que el planteamiento tradicional de la tarea.

Nuestras propuestas se han probado en un diseño experimental común que usa tres conjuntos de datos de recomendación de películas y música y cuatro algoritmos de recomendación de referencia ampliamente conocidos. Los resultados de nuestros experimentos respaldan la validez de nuestras contribuciones y permiten el análisis y el entendimiento de su comportamiento cuando se aplican en diferentes configuraciones.

ACKNOWLEDGMENTS

> *"We all change. When you think about it, we're all different people all through our lives, and that's okay, that's good, you gotta keep moving, so long as you remember all the people that you used to be. I will not forget one line of this. Not one day. I swear. I will always remember when the Doctor was me."*
>
> The Doctor (Doctor Who, The Time of the Doctor)

There have been many people that have contributed to this thesis one way or another. Let these lines serve as a sign of my gratitude to all of them.

In the first place, I would like to thank my supervisor Pablo Castells for offering me the opportunity of pursuing this PhD. It was June of 2010, I was studying the last year of my Bachelor's degree as an exchange student in Bielefeld University (Germany) when I received his kind offer to join him and the rest of the Information Retrieval Group at the Universidad Autónoma de Madrid as a postgraduate student. At that time I was taking an introductory course to Information Retrieval taught by Prof. Philipp Cimiano, and the idea of doing a PhD on a related topic seemed back then a good idea...and turns out I was right! Four years and a half later, here I am writing these lines and proudly looking back at this period. Doing research is one of the most enriching activities I have done in my life, but it is also a hard and demanding endeavor that, without proper guidance, can lead to frustration and discouragement. Pablo's efforts and continuous work during this time have proven to be decisive for me to find the path for which this thesis is but a first step.

The rest of the members of the Information Retrieval Group deserve a special mention. Our chats and discussions have proven to be an invaluable source of knowledge that no book or paper can provide. My gratitude goes therefore to Fernando Díez, Iván Cantador, David Vallet, Sergio López, Alejandro Bellogín, Pedro Campos, Ignacio Fernández and all the interns, students and colleagues with whom I have shared space and time in the laboratory B-408.

I would also like to thank my collaborators during the three internships that I have done in the last years. First, I want to express my gratitude to Rodrygo. L.T. Santos, Craig Macdonald and Iadh Ounis, who kindly welcomed me as a visitor in the University of Glasgow and made space in their schedules to collaborate with me during the summer of 2012. Second, I would like to thank Linas Baltrunas and Alexandros Karatzoglou for giving me the opportunity to work with them as a summer intern in Telefónica Research Barcelona in 2013. Finally, I thank Roi

Blanco and Peter Mika for giving me the chance of being an intern at Yahoo! Research Barcelona this last summer.

Ninguno de mis pequeños logros en mi incipiente carrera investigadora hubiera sido posible sin la base adecuada. Afortunadamente, he tenido el privilegio de tener unos profesores magníficos en las distintas etapas educativas por las que he transitado. Mi agradecimiento en este sentido va hacia el personal docente e investigador del Departamento de Ingeniería Informática de Escuela Politécnica Superior y del Departamento de Matemáticas de la Facultad de Ciencias de la Universidad Autónoma de Madrid. Guardo un recuerdo muy especial de mis dos años de bachillerato en Colegio San Alberto Magno de Zaragoza, especialmente de Gerardo Alós y Miguel Ángel Per como profesores de Matemáticas y Física, respectivamente. Por último, quiero hacer un reconocimiento al Colegio Público San Braulio, concretamente a la labor de Javier Allué y Pilar Campo, cuya pasión por la enseñanza explican en buena medida mi trayectoria académica.

Pasando ya al plano personal, quiero hacer mención a esas personas indispensables que he conocido a lo largo de mi vida y que puedo presumir de llamar amigos. Por aguantarme en el día a día en estos últimos años, quiero agradecer a Gema y a Héctor Menéndez su compañía y apoyo. Junto con este último, no me podría olvidar de los mejores embajadores de Extremadura, José Luis y Luis Miguel, con quienes pasé largas jornadas de trabajo y estudio durante cinco inolvidables años de carrera. Por último, quiero mencionar con especial afecto a Héctor Calvo y Philipp, compañeros de memorables andanzas extra-académicas.

Finalmente, no puedo dejar de mencionar a los míos. Quiero mandar mi agradecimiento allá arriba a mis abuelos Carlos, Aurora, Félix y Dionisia por haber dado origen a mi extensa familia sin la cual yo no sería quien soy. En cuanto a mis padres Carlos y María Pilar, no existen palabras para reflejar mi gratitud por el amor y la fe que han depositado en mí. Nunca han escatimado tiempo y recursos para que tanto mi hermana Sheila como yo recibiéramos la mejor educación posible. A Sheila quiero agradecerle el haber estado siempre ahí y, muy especialmente, el tiempo que hemos pasado estos años perdidos en la inmensidad de Madrid.

Saúl Vargas Sandoval
February 2015

CONTENTS

LIST OF FIGURES

LIST OF TABLES

1

INTRODUCTION

1.1 MOTIVATION

Recommender Systems have become a pervasive technology in a wide spectrum of everyday applications, and can be said to be familiar to the general public. Since the creation of the World Wide Web in 1991 (Berners-Lee, 1992), the amount of content and resources in it has risen exponentially. The case of e-commerce, streaming platforms and social networks, which constitute nowadays a substantial part of the Web's traffic, is specially interesting. Typically, such services offer a varied and vast amount of content to their customers: more than 200 million products in Amazon.com, 30 million songs in Spotify, 10,000 movies in Netflix, 248 million active users in Twitter sending 500 million messages everyday, etc. In such situations where there is an *information overload* (Toffler, 1970), assisting the users in exploring and finding resources of their interest is vital for the viability of such business models. Recommendation technologies, by means of providing personalized suggestions in such vast catalogs, have proven to be a major source of revenue and user satisfaction.

Recommender Systems have attracted an increasing level of interest in the academic community in the last two decades. This has resulted in an abundance of recommendation algorithms, technologies and software. Research on the area has been covered in many fora. The ACM Conference on Recommender Systems[1], which started in 2007, can be considered the main forum in the field. Recommender Systems are also a recurrent topic in other top tier conferences in Computer Science, such as WWW (Sarwar et al., 2001), SIGIR (Herlocker et al., 1999), CIKM (Karatzoglou et al., 2012), WSDM (Zhang et al., 2012), ICML (Salakhutdinov et al., 2007) or KDD (Niemann and Wolpers, 2013), to name a few. Journals such as TKDE (Yu et al., 2004), TOIS (Herlocker et al., 2004), IPM (Sweeney et al., 2008) or IRJ (Wang et al., 2008) are also a main source of published research in the field.

Despite the considerable progress in the area of Recommender Systems in the last two decades, there is a general awareness that there are still many open challenges and controversial issues that affect the current state of recommendation technologies and require further efforts. We identify novelty and diversity in Recommender Systems, which have attracted the attention of the community in the last decade, as fundamental quality dimensions in Recommender Systems whose

[1] http://recsys.acm.org/

study constitutes a promising direction for the advancement in the field. The contributions of this thesis are framed in this particular topic.

Novelty and diversity cover a set of different but interrelated perspectives that affect the quality of recommendations in terms of user satisfaction and business performance. There are many situations where user satisfaction with recommendations can be enhanced by applying some degree of novelty or diversity. Consider the case of recommending the latest *summer hit*. Assuming that the user who receives such recommendation frequently listens to music, there is a high chance that she already knows it. In this case, suggesting more novel (in the sense of less popular) songs might contribute to the utility of recommendations as tools for discovery of new, unknown content. In a different setting, recommending a list of movies consisting only of, say, westerns, however relevant to the user, can be highly redundant and unsatisfactory to the user's needs. Users tend to have a variety of interests and tastes and a need or desire for varied recommendations, so matching only one particular movie genre might result in a sub-optimal recommendation. In terms of business performance, novelty and diversity can also be considered. For example, providing different recommendations to different users makes sense from a business perspective: not only the users are interested in exploring the catalog but also the business is interested in making the whole catalog visible to the users. A system that provides highly relevant recommendations using only one tenth of the catalog might satisfy the users, but is sub-optimal from a business point of view. These and other perspectives on novelty and diversity in recommendations are the object of this thesis.

The phrase *"If you cannot measure it, you cannot improve it"*, attributed to Lord Kelvin, applies perfectly to the evaluation of Recommender Systems. Indeed, properly assessing the performance of recommendations it terms of the different quality dimensions involved is the first step towards making them useful. Since the beginning of the 2000's, an increasing stream of proposals has resulted in a variety of metrics for the assessment of the different perspectives on novelty and diversity in Recommender Systems. However, we find such set of metrics to be highly heterogeneous and lacking a detailed analysis about the difference, equivalences and connections between them. Furthermore, these metrics also lack, in many cases, properties as important as considering the real utility provided to the user in terms of the rank and relevance of the recommended items (movies, music, book or other types of products): no matter how novel is an item in a recommendation, very little utility is obtained from it if the user dislikes it or does not even see it. We consider that we need to revisit the related work in this topic under a renovated perspective. In particular, we find in the degree of formalization of Information Retrieval evaluation, especially in what concerns the assessment of the diversity of search results, a promising source of theories and concepts that might help lay out new views on the evaluation of recommendations.

The area of Information Retrieval addresses the broader task of providing the users with easy access to information of their interest. It deals with the representation, storage, organization of, and access to information items such as documents, web pages, online catalogs, structured and semi-structured records and multimedia objects (Baeza-Yates and Ribeiro-Neto, 2011). Web search engines are the most visible Information Retrieval applications. Given a user who expresses some information need in the form a (short) query, the task of a Web search engine consists in returning a search result composed of web pages relevant to the issued query. Recommender Systems can be viewed as a special case of an Information Retrieval system, in which the information need is expressed implicitly – that is, a query is generally absent – and therefore the personalization is particularly decisive for satisfying the user's information need. It is thus natural to contemplate common views between search and recommendation and adapt techniques from one field to the other. An important part of our contributions results from adapting notions from Information Retrieval to Recommender Systems.

The evaluation of Information Retrieval systems has been characterized by the formalization of metrics and evaluation methodologies under well understood concepts and elaborate user models (e. g. Carterette (2011)). We believe such level of rigor can benefit the still incipient evaluation of novelty and diversity in Recommender Systems. Moreover, diversity in Information Retrieval also plays an important role to cope with query ambiguity and underspecification. A query such as "java" could refer to the programming language or the Indonesian island. In this case, presenting documents covering these and other possible interpretations is an effective strategy to satisfy the possible underlying information needs behind the query. In the spirit of looking for common perspectives between search and recommendation, we find the motivation for search diversification to be applicable to recommendation: users tend to have a variety of interests that, in the absence of any other information, needs to be addressed in the recommendations they receive. Therefore, we think that exploring the adaptation of search result diversification to the recommendation task can lead to further benefits in the evaluation and enhancement of novelty and diversity in Recommender Systems.

Recommender Systems, however, present particularities of their own that need to be addressed considering domain-specific motivations and techniques. The diversity within recommendations does not only solve a potential problem of ambiguity of users' needs, but also addresses the need or desire for varied recommendations. Therefore, a specific analysis of the properties of diverse recommendations is required to go beyond what the state of the art in recommendation and search result diversification techniques offer. The diversity among recommendations delivered to different users is also a specific problem barely addressed in Information Retrieval. Most recommendation scenarios present a so-called *long tail effect* (Anderson, 2006), in which a few of the most popular resources (the short head) account for a significant portion of the interactions with users (views, ratings, sales), as opposed to the rest (the long tail). It has been argued that a recommender system that promotes recommendations in the long tail not only provides benefits for the business in terms of making the most of the catalog, but also help providing the user less known and obvious recommendations (Celma and Herrera, 2008). In this thesis, we address such recommendation-specific problems and propose solutions for their enhancement.

1.2 RESEARCH GOALS

The general aim of this thesis is to propose a principled approach to the evaluation and enhancement of novelty and diversity in Recommender Systems. We consider that the improvement of such fundamental dimensions of the usefulness of recommendations has to take into account how users explore and perceive recommendations, what are the problems that novelty and diversity solve, and the causes of such problems. To do this, we have pursued the following research goals.

RG1: revisit the work in novelty and diversity evaluation for recommendations and develop a clear common methodological and conceptual ground that takes into account how users perceive the utility of recommendations. As stated in the motivation, we find in the related work on evaluation of novelty and diversity in Recommender Systems a wide set of metrics, but a clear understanding of the connections and differences between the perspectives behind the metrics is missing in the literature. Furthermore, many of the metrics lack important properties reflecting how users interact with recommendations and the utility these provide.

RG2: explore the application of theories and methods from Information Retrieval diversity to Recommender Systems. By linking the recommendation and search tasks, we investigate the benefits of applying search result diversification techniques to the recommendation task. In particular, we establish an analogy between the ambiguity and underspecification of short queries submitted in search engines and the ambiguity of users' interests and tastes, plus the inherent uncertainty involved in the evidence of such interests available to a recommender system.

RG3: devise recommendation-specific techniques for enhancing the diversity within recommendations. Beyond bridging perspectives between Information Retrieval and Recommender Systems, the need for varied recommendations lies outside the goal of search result diversification, and requires dedicated approaches in a recommendation setting.

RG4: proposing new techniques for alleviating the popularity bias in recommendations. As we mention in the motivation, one of the main causes for the lack of novelty and diversity is the bias in recommendation algorithms towards highly popular items. We propose new methods to alleviate such effect while maintaining the accuracy of recommendations.

1.3 CONTRIBUTIONS

The work carried out throughout this thesis has resulted in several contributions to the research in novelty and diversity in Recommender Systems, which we summarize next.

In **Chapter** 4 we propose a **unified framework for novelty and diversity in Recommender Systems**. This framework is based on three fundamental relations between users and items, namely discovery, relevance and choice, and two configurable components, namely an item novelty model and a browsing model, which together define and generalize many of the metrics for the different notions of novelty and diversity in the state of the art. Apart from providing a formal ground for metrics of the state of the art, our framework supports additional properties such as rank and relevance awareness in recommendation lists.

Chapter 5 proposes the **diversification of recommendations by means of adapting the family of Intent-Aware metrics and diversification methods of search result diversification**. For this adaptation, an analogy between query interpretations or facets and user interests or tastes is established. This allows us to adapt metrics and diversification methods of search result diversification to the recommendation task. In this context, we propose two new methods to enhance the diversity of recommendations. The first method introduces an explicit relevance model that replaces the generative model found in the adapted diversification techniques for search result diversification. The second method makes use of sub-profiles to provide recommendations suited to each different taste or interest of a user, which are later combined into a single, diversified recommendation.

We study in **Chapter** 6 the specific case of modeling diversity in recommendations when items are categorized by means of genres, as is the case of movies, music or books. We identify the requirements for genre-based diversity, namely coverage, redundancy and size-awareness, and argue that none of the previous diversification frameworks satisfy them. We propose a new **Binomial framework for modeling genre-based diversity** that satisfies these requirements.

Finally, **Chapter** 7 proposes two methods to alleviate the effect of the popularity-biased concentration in collaborative filtering recommendations. Both approaches result from turning the recommendation task around by conceptually **recommending users to items**. The first method is a new policy to select neighbors in nearest neighbors algorithms. The second method develops a probabilistic reformulation of the recommendation task that isolates and controls the effect of the popularity bias in recommendations.

1.4 PUBLICATIONS

The work carried out for the completion of this thesis has resulted in various publications in international conferences, journals, book chapters and other fora. We list these publications next, sorting them by publication type and the chapter they are related to.

Publications Related to Chapter 2

Book chapters:

- Castells, P., Hurley, N., and Vargas, S. (in press). Novelty and diversity in recommender systems. In Ricci, F., Rokach, L., and Shapira, B., editors, *Recommender Systems Handbook 2nd Edition*. Springer US

Publications Related to Chapter 4

Long papers in international conferences and journals:

- Vargas, S. and Castells, P. (2011). Rank and relevance in novelty and diversity metrics for recommender systems. In *Proceedings of the 5th ACM Conference on Recommender Systems*, RecSys '11, pages 109–116, New York, NY, USA. ACM

Workshop papers, national publications and posters:

- Castells, P., Vargas, S., and Wang, J. (2011). Novelty and diversity metrics for recommender systems: Choice, discovery and relevance. In *International Workshop on Diversity in Document Retrieval at the 33rd European Conference on Information Retrieval*, DDR'11

Publications Related to Chapter 5

Long papers in international conferences and journals:

- Vargas, S., Castells, P., and Vallet, D. (2012a). Explicit relevance models in intent-oriented information retrieval diversification. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '12, pages 75–84, New York, NY, USA. ACM

- Vargas, S. and Castells, P. (2013). Exploiting the diversity of user preferences for recommendation. In *Proceedings of the 10th Conference on Open Research Areas in Information Retrieval*, OAIR '13, pages 129–136, Paris, France. CID
- Vargas, S., Santos, R. L. T., Macdonald, C., and Ounis, I. (2013). Selecting effective expansion terms for diversity. In *Proceedings of the 10th Conference on Open Research Areas in Information Retrieval*, OAIR '13, pages 69–76, Paris, France. CID

Workshop papers, national publications and posters:

- Vargas, S., Castells, P., and Vallet, D. (2011). Intent-oriented diversity in recommender systems. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '11, pages 1211–1212, New York, NY, USA. ACM
- Vargas, S. and Castells, P. (2012). Diversificación en sistemas de recomendación a partir de sub-perfiles de usuario. In *II Congreso Español de Recuperación de Información*, CERI'12

Publications Related to Chapter 6

Long papers in international conferences and journals:

- Vargas, S. and Castells, P. (2014a). Improving sales diversity by recommending users to items. In *Proceedings of the 8th ACM Conference on Recommender Systems*, RecSys '14, pages 145–152, New York, NY, USA. ACM

Workshop papers, national publications and posters:

- Vargas, S., Castells, P., and Vallet, D. (2012b). On the suitability of intent spaces for IR diversification. In *Proceedings of the International Workshop on Diversity in Document Retrieval at the 5th ACM International Conference on Web Search and Data Mining*, DDR'12, Seattle, Washington, USA

Publications Related to Chapter 7

Long papers in international conferences and journals:

- Vargas, S., Baltrunas, L., Karatzoglou, A., and Castells, P. (2014). Coverage, redundancy and size-awareness in genre diversity for recommender systems. In *Proceedings of the 8th ACM Conference on Recommender Systems*, RecSys '14, pages 209–216, New York, NY, USA. ACM

Workshop papers, national publications and posters:

- Vargas, S. and Castells, P. (2014b). Vecindarios inversos para la mejora de la novedad en filtrado colaborativo. In *III Congreso Español de Recuperación de Información*, CERI'14

Other Publications Related to the Thesis

Presentation of this thesis in specialized doctoral symposia and consortia:

- Vargas, S. (2011). New approaches to diversity and novelty in recommender systems. In *Proceedings of the 4th BCS-IRSG Conference on Future Directions in Information Access*, FDIA'11, pages 8–13, Swinton, UK. British Computer Society

- Vargas, S. (2014). Novelty and diversity enhancement and evaluation in recommender systems and information retrieval. In *Proceedings of the 37th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '14, pages 1281–1281, New York, NY, USA. ACM

1.5 STRUCTURE OF THE THESIS

The thesis is structured as follows:

- Chapter 1 introduces this thesis by presenting the motivation, research goals, contributions, publications related to the thesis and the definitions and notation.

- Chapter 2 reviews the related work on the topics of interest for this thesis. First, we provide a general overview of the state of the art in Recommender Systems. Second, we delve into the study of novelty and diversity in Recommender Systems. Last, we examine the work on search result diversification in Information Retrieval.

- Chapter 3 presents the design of the experiments of our contributions. In particular, we provide a detailed description of the datasets, recommendation algorithms and evaluation methodology that we have used in the different experimental sections of our contributions of the following chapters.

- Chapter 4 proposes a unified framework for novelty and diversity in Recommender Systems that contributes to the formalization of metrics and reranking techniques and the consideration of properties such as rank and relevance in recommendations.

- Chapter 5 presents an adaptation of the Intent-Aware metrics and diversification methods in search result diversification to Recommender Systems. On top of this adaptation, we provide two new methods that improve over direct adaptations of the diversification methods of search result diversification.

- Chapter 6 addresses the problem of assessing and optimizing Intra-List Diversity when genre information about the items in a domain is used. Three requirements for genre diversity are identified: coverage, redundancy and recommendation list size-awareness. We propose a Binomial framework that satisfies these three requirements and compare it with related approaches to measure Intra-List Diversity.

- Chapter 7 presents our contributions for the improvement of Sales Diversity. By conceptually recommending users to items, we present two different approaches, namely inverted nearest neighborhoods and a probabilistic reformulation layer, that offer significant improvements in terms of Sales Diversity when compared with prior proposals.

- Chapter 8 offers the conclusion and future work.

- Appendix A contains the high-level documentation for RankSys, a new recommendation framework developed for this thesis that specializes in novelty and diversity evaluation and enhancement in Recommender Systems.

- Appendix B contains the translation into Spanish of Chapter 1.

- Appendix C contains the translation into Spanish of Chapter 8.

1.6 DEFINITIONS AND NOTATION

We summarize here for the reader's convenience the main definitions and notation that we shall use all over the rest of this thesis. Additional specific notation, when necessary, will be described in the chapter where it applies.

Without loss of generality, the recommendation problem can be formulated as suggesting items from a catalog $\mathcal{I}$ in a particular recommendation scenario – products, movies, music or other resources – to a community of users $\mathcal{U}$. In order to make personalized recommendations, we require some previous knowledge about the users in the form of the items they have rated, consumed, bought, etc. This interaction data, that we denote as $\mathcal{R}$, is typically encoded in the form of a $|\mathcal{U}| \times |\mathcal{I}|$ matrix whose elements r_{ui} represent the interaction between users and items. In its simplest form, we can consider this interaction matrix to take values 0 and 1, i. e. $\mathcal{R} \in \{0, 1\}^{|\mathcal{U}| \times |\mathcal{I}|}$, when we only have binary interactions between users and items: the user bought, consumed, watched or listened to the item, etc. In the case when users provide numerical ratings, for example 1 to 5 stars, $\mathcal{R}$ takes values in $\{0, 1, 2, 3, 4, 5\}$ where by 0 we denote the absence of a rating. In settings where we can quantify the interaction between users and items, for example play counts of music, our interaction matrix can take values in the set of natural numbers $\mathbb{N}$. More complicated cases, such as time data for the interactions, can be easily accommodated with similar formulations. Conveniently, by abuse of notation we can re-interpret, for all previous cases, the interaction matrix $\mathcal{R}$ as the pairs of users and items $\mathcal{R} \subset \mathcal{U} \times \mathcal{I}$ that had any kind of interaction. Under this interpretation, we denote by $\mathcal{I}_u = \{i \in \mathcal{I} : (u, i) \in \mathcal{R}\}$ the user profile or subset of items that user u interacted with, and, respectively, by $\mathcal{U}_i = \{u \in \mathcal{U} : (u, i) \in \mathcal{R}\}$ the item profile or subset of users we know that had any interaction with item i.

Using the interaction data $\mathcal{R}$, the goal of a recommender system S consists in generating recommendations R_u^S for every user u. A recommendation is a set of items $R_u^S \subset \mathcal{I}$ that are presented to the user. Frequently, recommendations are presented as a ranked list, so we may also interpret them as a sequence of items $R_u^S \in \mathcal{I} \times \ldots \times \mathcal{I}$. In our setting, recommended items are selected by means of a scoring function $s : \mathcal{U} \times \mathcal{I} \to \mathbb{R}$ that determines the ranking R_u^S of the recommended items in decreasing order of the scores, that is, if $s(u, i) \geqslant s(u, j)$ then $R_u^S = (\ldots, i, \ldots, j, \ldots)$. In the remainder of this document we will drop for convenience the user or the system in the notation of a recommendation R_u^S and leave it as R for short, except when we need to explicitly indicate the user or the system.

Along with the interaction data $\mathcal{R}$, certain metrics may employ additional information about the items when assessing some novelty and diversity perspectives. Generically, we denote as $\mathcal{F}$ a set of features or characteristics about the items, and as $\mathcal{F}_i$ the subset of features that describe an item i. Examples of such features depend on the type of the items. For instance, in the case of movies or songs, we could consider language, year of release, genre, etc. As we describe in Chapter 3, we consider in our experimental design the specific case of genres for movie and music recommendation. In that case, we denote as $\mathcal{G}$ the set of genres in a specific recommendation domain and, for an item i, we denote as $\mathcal{G}_i$ the genres covered by the item.

2

RELATED WORK

2.1 INTRODUCTION

As a fundamental part on the work carried out for this thesis, a study of the state of the art in the areas of Recommender Systems and Information Retrieval has been conducted. In particular, we have been interested in the revision of the work related to novelty and diversity in Recommender Systems, without losing a broader perspective of the latest advances in the area, specially in evaluation methodologies and collaborative filtering approaches. Also, as part of our vision of the recommendation problem as an Information Retrieval problem, we see the study of search result diversification as a potential source of new perspectives and methods for the diversity of recommendations.

In this chapter we present a review of the related work on the topics of interest of this thesis. First, a global overview of the Recommender Systems area is presented in Section 2.2. Then, in Section 2.3 we delve into the study of the related work on Novelty and Diversity in Recommender Systems. Finally, Section 2.4 covers the literature in search result diversification for Information Retrieval.

The contents of this chapter are partly available in the following published work:

- Castells, P., Hurley, N., and Vargas, S. (in press). Novelty and diversity in recommender systems. In Ricci, F., Rokach, L., and Shapira, B., editors, *Recommender Systems Handbook 2nd Edition*. Springer US

2.2 RECOMMENDER SYSTEMS: AN OVERVIEW

Recommender Systems (see (Resnick and Varian, 1997; Adomavicius and Tuzhilin, 2005; Ricci et al., 2011)) are software tools and techniques that provide suggestions for users about a catalog of items – such as products, video, music, or other resources – in a personalized manner. Recommender Systems become specially useful in scenarios where there is an *information overload*, that is, the overwhelming array of choices makes the exploration and selection in the catalog a difficult task for the user. The personalized assistance in the exploration and discovery of content that recommender systems offer has proven to be an effective way of increasing user satisfaction and improving revenue of many e-commerce and media streaming platforms such as Amazon, Netflix, Youtube or Spotify and social networks such as Facebook or Twitter.

Recommender systems build on a set of multidisciplinary theories, technologies and algorithms from varied fields such as Information Retrieval, Machine Learning, Human Computer Interaction, Marketing, Economics and many others. Recommender Systems remains an active topic of research that has attracted an increasing level of attention in the last decade.

In the remaining of the section, we review the state of the art in Recommender Systems. First, we focus on the particularities of the different scenarios where recommendations can be offered. Then, we provide a classification of the recommendation algorithms that are found in the literature. The evaluation of Recommender Systems, which is a central topic of this thesis, is also introduced here. Afterwards, we comment several of the issues, limitations and challenges that current research in the area faces. Finally, we present some of the tools and services available for the implementation and deployment of recommendation technologies.

2.2.1 *Recommendation Scenarios*

Recommender Systems can be used in multiple scenarios where there is an *information overload* and a need to assist the user in exploring a vast catalog and discover new, relevant resources. One of the most studied cases, probably because of the simplicity of the data, the genuine fit of the recommendation task in this domain, and the early availability of public datasets, is the movie recommendation scenario, for which MovieLens and Netflix are the typical examples. In this scenario, the recommendation task is posited as the suggestion of movies from a broad catalog so that they match the preferences of the users as expressed in the ratings they assign to movies they have watched. There are however many other recommendation scenarios in which some of the generally assumed conditions of movie recommendations do not hold. In particular, we propose a general characterization of the most common recommendation scenarios attending to the following set of criteria:

- Type of feedback: the preferences of the users may be expressed in an explicit or implicit manner.

- Single selection and repeatability: the user may be interested in selecting several recommended items at once or consume the same item (or a similar one) repeatedly over time.

- Scarcity and volatility: items may be short-lived and/or subject to stock availability.

- Context: recommendations may depend on short-term interests of the current session.

- Reciprocity: the recommended resources may actually be other users and, in this case, the relevance between target and recommended users must be mutual.

The type of feedback gathered from interactions between users and items largely determines the choice of algorithms and evaluation methodologies that can be applied to a particular recommendation scenario. User feedback can be classified into two categories: explicit, when the user consciously expresses a preference assessment for the items; and implicit, in which case the preference is indirectly estimated by other variables that do not require the active involvement of the user. On one side, in the case of explicit feedback, the user may manifest her preferences in varied degrees of detail: simple binary feedback (like/dislike), numerical ratings (1-to-5 stars, 1-to-10 points), multi-criteria ratings (Adomavicius and Kwon, 2007) and written reviews. The degree of detail of the explicit feedback defines the potential utility and complexity of the recommendation approaches: working with binary feedback is relatively easy but provides limited information and, on the contrary, written reviews provide a lot of potentially useful information but their processing may require from advanced natural language processing techniques. On the other side, implicit feedback provides an indirect means of determining the preferences of the user for the items as in the case of play counts for music or videos (Hu et al., 2008) or clicks in a ranking context (Hofmann et al., 2014). Implicit feedback is generally much easier to obtain than explicit feedback, since it does not require any additional input from the user beyond her natural interaction with the system. However, working with such indirect evidences introduces uncertainty in the recommendation process, since a number of assumptions needs to be made to convert data such as play counts or clicks into estimations of user preferences. This categorization of user feedback does not impose however a separation or incompatibility between different signals. In practice, in many recommendations scenarios different types of feedback may be available at once. Designing recommendation approaches that make the most of the different signals in a combined manner may provide additional benefits that could not otherwise be obtained in single-type feedback techniques.

The nature of the selection of items in a recommendation also establishes differences between recommendation scenarios. For instance, given a recommendation, the user may be interested in selecting a single item, or more than one. The first would be the case, for example, of a user interested in buying a washing machine: most users will buy, at most, a single item. An opposite case would be for instance *playlist* music recommendation (Coelho et al., 2013): given a set of songs, the user selects some (or even all) of them to be played. A related but different case would be the recurrent consumption of the same or similar items. Consider again the case of washing machine recommendations: once a user buys a washing machine, the probability that she may be interested in buying another washing machine just a few days later is low. On the contrary, a recommender system in the grocery domain may recurrently recommend to a user the same type of food weekly, since it can safely assume that the user consumes many products periodically.

Another set of characteristics that define several recommendation domains relates to the volatility and scarcity of items. The assumption that items in the recommendation domain may be of interest to the user independently from its recentness does not necessarily hold, for instance, in news recommendations: last week's news are generally less relevant to most users than today's news. Also, in many recommendation domains the recommended items are subject to limited availability, as in the case of hotel room recommendation (Cremonesi et al., 2013) or products with limited stock. Unlimited availability is normally exclusive of digital content, such as movies, music in streaming services and e-books.

A fourth axis is defined by the context in which recommendations are presented. In many recommendation scenarios, recommendations are driven by the specific short-term interests of the current browsing session of the user. That is the case of "you may also be interested in" or "frequently bought together" suggestions when browsing a product in Amazon or Zalando (Jannach et al., 2013) or related videos in services such as Youtube. The previous cases contrast with context-free recommendations, such as those that can be offered in the homepage of the previously mentioned services targeting other long-term interests of the user.

A final division considers the reciprocity between users and items. Consider the case of people recommendation in online dating services (Pizzato et al., 2010). In this case, reciprocity plays a determinant role: both the recommended and the target users must be mutually compatible, or the purpose of the recommendation may not be fulfilled. Reciprocity may also be found in job recommendations (Malinowski et al., 2006): candidates must be offered relevant job offers, but employers also require that their jobs offers are presented to adequate candidates. Reciprocity can also be considered in other people recommendation scenarios, such as contact recommendation in Social Networks. In the rest of the recommendation scenarios, where items are inanimate objects, the reciprocity has been naturally disregarded, although Said et al. (2014) suggested that it can be considered as a means to address scenarios where item related factors such as ephemerality, novelty and interestingness may limit the potential users the items can be recommended to.

2.2.2 *Classification of Recommendation Algorithms*

Recommendation algorithms can be classified according to multiple criteria, such as the type of feedback they employ, the type of recommendation task they solve, etc. However, the most common division found in the literature refers to how they use the information of user preferences with respect to items, for which three categories are commonly established:

- Collaborative Filtering algorithms: recommendations are solely based on the consumption patterns of the users.
- Content-based algorithms: recommendations are based on the content or features of the items in the recommendation domain.
- Hybrid algorithms: combination of the previous methods.

We proceed now to succinctly review each family of recommendation algorithms, discussing their advantages, weaknesses and the main proposals. We make a special emphasis on Collaborative Filtering approaches as our experiments in the following chapters rely on this family of algorithms.

2.2.2.1 *Collaborative Filtering*

Collaborative Filtering describes the family of algorithms that exploit the users' consumption patterns of the items in the recommendation domain, without making use of any domain-specific characteristics of the items, such as their content or categorization. The main advantage of this type of algorithms is their independence with respect to the recommendation domain in which they are applied. They have been claimed to be more effective than other approaches, such as the content-based algorithms. Collaborative Filtering algorithms can be themselves classified into two types:

- Memory-based: the interactions between users and items are directly used to generate recommendations.
- Model-based: the recommendations are based on a model that is previously learned from the user data.

Memory-based methods are characterized by their simplicity, since minimal or no learning phase is involved. This lack of learning phase provides several advantages, such as easiness of implementation, immediate incorporation of new data and comprehensibility of results. Memory-based methods, however, may suffer from scalability issues and lack of sensitivity to sparse data (Lemire and Maclachlan, 2005).

The best known memory-based approaches are the so-called neighbors methods, which are divided in user and item-based. In the **user-based** methods (Desrosiers and Karypis, 2011), similarities between users in their consumption patterns are used to compute recommendations. The idea is that, for a given user, the preferences of similar users, the neighbors, can serve as recommendations. Various approaches have been used to compute the similarity between users, distinguishing between those that consider the agreement rate of common ratings between users (Resnick et al., 1994; Shardanand and Maes, 1995) or, more recently, the co-occurrence of items in user profiles (Cremonesi et al., 2010; Aiolli, 2013). Once the similarities between users are assessed, user-based methods generate, in their simplest formulation, recommendations for a user u by scoring the items in the profiles of the neighbors as a sum over the preference values assigned by the neighbors weighted by the similarity to the target user:

$$s_{UB}(u,i) = \sum_{v \in N(u)} sim(u,v)\; r_{v,i}$$

where $sim(u,v)$ is the similarity value between users and $N(u)$ denotes the set of neighbors of user u. For both efficiency and effectiveness reasons, these neighborhoods are usually restricted to consider a reduced set of highly similar users, either by establishing a minimal similarity threshold value (Zadeh and Carlsson, 2013) or selecting the k most similar users (Desrosiers and Karypis, 2011). Many variations and extensions of the previous scheme have been proposed. For instance, in the rating prediction problem (see Section 2.2.3) the scores are normalized by the sum of similarities to provide scores in the same range of the ratings in the preference data $\mathcal{R}$. Furthermore, user ratings deviations are usually added to the scoring function to avoid biased estimations:

$$s_{UB}(u,i) = \hat{r}_u + \frac{1}{\sum_{v \in N(u)} sim(u,v)} \sum_{v \in N(u)} sim(u,v)\;\; (r_{v,i} - \hat{r}_v)$$

where $\hat{r}_u$ is the average rating of user u. Other approaches break the assumption that nearest neighbors provide the best recommendations. For example, in (Said et al., 2013) a "furthest neighbors" approach is presented to provide more diverse recommendations. In the spirit of alleviating the obviousness of recommendations, Adamopoulos and Tuzhilin (2014) proposed a probabilistic neighborhood selection in which neighbors are randomly selected according to a probability proportional to their similarity to the target user. In Chapter 7 we propose a method for the selection of inverted nearest neighbors to improve the diversity of sales.

In the **item-based** methods (Sarwar et al., 2001), the similarities between items with common users are exploited. The idea is that items that are similar to those that the user has already rated or consumed are good candidates candidates for recommendation. In its plainest variant, for a user u this algorithm scores items similar to those of her profile $\mathcal{I}_u$ as the sum of their item-to-item similarities weighted by the preferences of the target user:

$$s_{IB}(u,i) = \sum_{j \in \mathcal{I}_u} \mathrm{sim}(i,j)\, r_{u,j}$$

Note that in this case we can also consider limited size neighborhoods of the items in the user profile, that is, recommending only those items in $\bigcup_{j \in \mathcal{I}_u} N(j)$ where $N(j)$ is the neighborhood of item j. In the same line of the user-based algorithm, many variants to the previous formula have been proposed. The item-based algorithm has the advantage that item-to-item similarities can be easily pre-computed to efficiently generate recommendations for all users. For instance, Amazon is known to use this approach to provide relevant recommendations in real time for their massive datasets (Linden et al., 2003).

Other memory-based methods include the probabilistic framework of Yu et al. (2004), the application of associate retrieval techniques of Huang et al. (2004) or, more recently, the reformulation of Verstrepen and Goethals (2014) of nearest neighbors methods that unifies both user and item-based variants for the one-class collaborative filtering problem.

Model-based methods take a different approach to exploit collaborative filtering data. The algorithms of this family depend on a learning phase, in which a descriptive model of user preferences based on the observed data is built to make predictions. These methods are inspired in machine learning techniques such as artificial neural networks (Salakhutdinov et al., 2007), Bayesian networks (Breese et al., 1998), clustering (Ungar and Foster, 1998) and latent factor models (Blei et al., 2003; Hofmann, 2004; Koren et al., 2009). Among these approaches, latent factor models are the most studied and widespread model-based techniques. These techniques perform a dimensionality reduction of the user-item matrix $\mathcal{R}$ in which a set of latent variables is used to explain user preferences for recommendation purposes. Such techniques include matrix factorization (Koren et al., 2009), probabilistic Latent Semantic analysis (Hofmann, 2004) and Latent Dirichlet Allocation (Blei et al., 2003).

Matrix factorization is one of the best known approaches and consists in obtaining a two user and item matrices $P \in \mathbb{R}^{|\mathcal{U}|,k}$ and $Q \in \mathbb{R}^{|\mathcal{I}|,k}$ that represent all the users and items in a k-dimensional latent vector space, where k is typically much smaller than the number of users or items. Such user and item matrices are obtained by minimizing some error or loss function $\mathcal{L}(P,Q)$ with respect to the observations of a user-item matrix $\mathcal{R}$ by varied methods such as stochastic gradient descent (Rendle and Freudenthaler, 2014; Shi et al., 2012a), alternating least squares (Hu et al., 2008; Takács and Tikk, 2012) or maximum margin matrix factorization (Weimer et al., 2007). Once P and Q are computed, the recommendations are determined by the scores generated by multiplying user and item vectors:

$$s_{MF}(u,i) = P_u \cdot Q_i^t$$

where P_u is the row vector of P that corresponds to user u and Q_i the row vector of Q which describes item i in the latent vector space.

In Chapter 3 we give further details about the Collaborative Filtering algorithms that we used in the common experimental design of our contributions, namely user and item-based nearest neighbors, implicit Matrix Factorization and probabilistic Latent Semantic Analysis.

2.2.2.2 *Content-Based*

Content-based methods (Lops et al., 2011; Pazzani and Billsus, 2007) build user profiles based on the features and descriptions of the items rated by the user and, contrary to Collaborative Filtering, do not use other users' preferences for issuing recommendations. One of the advantages of content-based methods is that they can deal seamlessly with the *new item* problem, that is, they are able to recommend new items for which there is no user feedback, as opposed to collaborative filtering algorithms. Content-based algorithms, however, are very dependent on the recommendation domain, which contrasts with the generality of collaborative filtering methods. Another drawback is that they also rely on the availability of enough and accurate information about the features of the items, which is sometimes costly to obtain. Finally, content-based approaches may (not exclusively but particularly) suffer from over-specialization, that is, they have a natural tendency to recommend items that are *too* similar to items that the user has already rated.

Proposals for content-based recommendation algorithms draw perspectives and algorithms from varied fields such as Information Retrieval, Semantic Web, and Machine Learning. For example, term-weighting models from Information Retrieval were used in early proposals for Web recommendations (Balabanović and Shoham, 1997), news recommendation (Lang, 1995), and, more recently, to social tagging systems in (Cantador et al., 2010). Approaches with Semantic Web technologies have also been proposed for content-based recommendations, as in the case of news recommendation (Cantador et al., 2008), or movie and music recommendations leveraging Linked Open Data (Ostuni et al., 2013). Regarding the use of Machine Learning techniques, Mooney and Roy (2000) used Bayesian classifiers for book recommendations and Pazzani and Billsus (1997) used various techniques such as Bayesian classifiers, clustering, decision trees and artificial neural networks for Web site recommendation.

2.2.2.3 *Hybrid approaches*

Hybrid methods (Burke, 2002; Adomavicius and Tuzhilin, 2005) have been proposed to avoid the limitations of collaborative filtering and content-based algorithms when used separately. As identified by Adomavicius and Tuzhilin (2005), the main trends of hybrid recommendation approaches can be classified as follows:

- Combining separate recommendations: the predictions of separate recommendation algorithms are combined to provide a single recommendation, using methods such as linear combinations (Claypool et al., 1999) or voting schemes (Pazzani, 1999).
- Adding content-based characteristics to collaborative filtering: Pazzani (1999) adapted the user-based neighbors method to calculate similarities based on content-based user profiles.
- Adding collaborative characteristics to content-based methods: latent factor models can be applied to content-based approaches for text recommendation, as in (Soboroff and Nicholas, 1999).
- Developing a single unifying recommendation model: in the work of Popescul et al. (2001) and Schein et al. (2002) a unified probabilistic method for combining collaborative and content-based recommendations is presented.

2.2.3 *Evaluation Methodologies*

The evaluation of Recommender Systems (Shani and Gunawardana, 2011) plays a significant role in the development of new proposals, and still constitutes an active topic of research. Given the complexity and the amount of factors involved in a recommendation setting, the evaluation of Recommender Systems must be assessed from different points of view. We differentiate three main axes in which we can classify the existing evaluation methodologies:

- offline or online: the evaluation is performed with collected information about users, or evaluated in real time as users interact with the system.
- user vs. business-oriented: some measurements attempt to quantify the user satisfaction with respect the system while others measure directly the variables directly affecting the business or platform running the recommender system under evaluation.
- accuracy or alternative measurements: typically, evaluation has been oriented towards assessing the capacity of the recommendations to provide items that are relevant to the user, although there are many other quality dimensions involved in a successful recommendation.

In the remaining of the section, we go into detail about this classification by reviewing the most common evaluation methodologies that can be found in the related work, emphasizing their particularities, weaknesses and strengths.

2.2.3.1 *Offline and Online Evaluation*

Offline Experiments

Offline experiments have been extensively used in the literature for validating recommendation algorithms since they allow a simple, cost-effective assessment of the performance of recommendation algorithms and the comparison between alternatives. An offline evaluation provides an *objective* way of measuring the performance of recommendation algorithms as they are based on metrics that capture one or more dimension qualities that do not depend on external conditions that may affect online evaluations. Offline evaluations depend on collected datasets that provide information about the tastes of the users in a particular recommendation domain.

Several recommendation datasets for varied recommendation domains are publicly available for research purposes. Some of the most popular datasets are the movie recommendation datasets of MovieLens[1]. In the context of the Netflix Prize[2], Netflix released a large-scale dataset to promote the research in the rating prediction problem of the challenge. Unfortunately, this dataset is no longer public due to anonymity concerns. For music recommendation, Ò. Celma collected two datasets from Last.fm[3] and, more recently, McFee et al. (2012) released user data for the Million Song Dataset[4]. In Chapter 3 we provide further details about the MovieLens1M, Netflix and Million Song datasets, which are the basis of our experiments in Chapters 4, 5, 6 and 7.

For a given a dataset consisting in user item interactions $\mathcal{R}$, the offline evaluation starts typically with the partition of the data into a training and test subsets: $\mathcal{R} = \mathcal{R}_{\mathtt{train}} \uplus \mathcal{R}_{\mathtt{test}}$. On the one hand, the training subset comprises the knowledge used for generating recommendations. On the other hand, the test subset is used as a proxy for determining the relevance for the users of the recommendations generated with the training data. A variety of splitting approaches have been proposed for creating partitions on the datasets. Such proposals can be classified into two major groups: random splitting (Goldberg et al., 2001; Sarwar et al., 2001), in which observations of interactions between users and items are randomly selected for train or test, and time-based splitting (Campos et al., 2014; Gunawardana and Shani, 2009), in which more recent interactions of the users in the recommendation domain are selected for testing and the older information is left for training purposes.

Once a partition for training and test has been performed, the literature distinguishes between two problems (Steck, 2013): the rating prediction and the ranking problem. The rating prediction problem, popularized with the Netflix Prize

[1] http://grouplens.org/datasets/
[2] http://netflixprize.com
[3] http://ocelma.net/MusicRecommendationDataset/
[4] http://labrosa.ee.columbia.edu/millionsong/

challenge, has been for many years the de-facto evaluation methodology in Recommender Systems. Such dominance has been fading in the last decade, favoring instead ranking-based methodologies that correspond better with real effectiveness.

Rating Prediction: The **rating prediction** problem can be stated as follows. The interactions between users and items are encoded in the form of a rating matrix $\mathcal{R}$ that contains partial information about the tastes of the users $\mathcal{U}$ for the items $\mathcal{I}$ in the form of discrete grades (such as 1 to 5 stars) that range from total dislike (1 star) to maximum enjoyment (5 stars). The goal of a recommender system S consists here in providing predictions for the ratings in the test subset based on the ratings present in the training subset. The evaluation of these predictions is normally based on error metrics, such as the mean-absolute Error (MAE) or the root-mean-square error (RMSE):

$$\mathrm{MAE}(S) = \frac{1}{|\mathcal{R}_{\mathrm{test}}|} \sum_{(u,i) \in \mathcal{R}_{\mathrm{test}}} |r_{u,i} - s(u,i)|$$

$$\mathrm{RMSE}(S) = \sqrt{\frac{1}{|\mathcal{R}_{\mathrm{test}}|} \sum_{(u,i) \in \mathcal{R}_{\mathrm{test}}} (r_{u,i} - s(u,i))^2}$$

As previously commented, rating prediction has been considered as the standard evaluation methodology for recommendation algorithms until recently. Such predominance can be explained by the availability of datasets for this task – many popular datasets, such as MovieLens, provide rating data – and the notoriety of the Netflix Prize, which awarded $1M to the team that proposed an algorithm that improved at least a 10% over the accuracy of the "Cinematch" algorithm of Netflix as measured by RMSE. There is however an increasing awareness that rating prediction methodologies may be inappropriate for the evaluation of Recommender Systems, as improvements of rating prediction may not correspond to actual improvements in user satisfaction of effectiveness of recommendations. In particular, rating prediction is founded on the assumption that missing ratings are *missing at random* (Marlin and Zemel, 2009), when the reality is the contrary: in real-world scenarios, users are free to rate the items they choose, and it has been observed that these choice patterns are not random. In fact, experimental evidence shows that rating choice patterns, as implicit as they are, are more informative than the values of the ratings. Moreover, as stated in Section 2.2.1, in many practical recommendations scenarios ratings are not available or, even when they are, they are accompanied by additional sources of feedback about the preferences of the users. The utility of the rating prediction task is therefore restricted to a specific type of feedback and, furthermore, it raises theoretical concerns because it ignores the non-random nature of users' rating patterns.

In contrast to rating prediction, the **ranking** task focuses on the more natural scenario of generating a ranked lists of items to present to the users for their evaluation. This approach mimics better many usual recommendation scenarios, in which the user expects the system to make a selection – a cut-off of the ranking – of the items in the catalog. This particular sub-task of selecting a cut-off of the N first ranked results is commonly known in the literature as the **top-N** recommendation task (Cremonesi et al., 2010).

Ranking task: The ranking task is usually formulated as follows. Given a user u and a set of candidate items, the task of a recommender system consists in producing a ranking R_u of candidate items according to the predicted relevance for the user. Generally, the ranking is determined by the scores $s(u, i)$ provided for the items by the recommender system in decreasing order. A variety of strategies for selecting the candidates items has been proposed in the literature. For example, several authors consider as candidate items only those in the test subset of the target user (Weimer et al., 2007; Shi et al., 2010; Niemann and Wolpers, 2013). This approach, however, may suffer from the same theoretical pitfalls as the rating prediction problem, namely it ignores the fact that missing ratings are *not missing at random* and, therefore, a system that produces a good ranking of the user test does not necessarily need to be optimal in ordering randomly chosen data. To overcome this problem, other methods consider the ranking of the items of the user test together with a selection of the rest of unobserved items at training time (Cremonesi et al., 2010; Bellogín et al., 2011), or all of them (Aiolli, 2013). The previous methods assume that unobserved ratings in either training and test time are irrelevant – which is a reasonable assumption given the observations of Steck (2013). Independently of the choice of candidate items, the assessment of the quality of rankings is generally measured by rank-aware metrics from Information Retrieval, such as precision or recall (Bellogín et al., 2011) or the normalized Discount Cumulative Gain (nDCG) of Järvelin and Kekäläinen (2000).

The ranking task, as opposed to rating prediction, is not limited to rating data as rankings of items may be generated by any type of explicit or implicit feedback. Moreover, rankings (and the sub-case of top-N recommendations) can be evaluated by more criteria than their accuracy. For instance, additional quality dimensions, such as novelty and diversity, become meaningful in the context of recommendation lists.

The difference between the rating prediction and ranking tasks can be further explained in the context of the *Learning to Rank* field, which considers the application of Machine Learning techniques to the document ranking problem in Information Retrieval (Liu, 2009) and, more recently, to the ranking task in Recommender Systems (Karatzoglou et al., 2013). This field of study as has emerged as an alternative to classical Machine Learning techniques, which focus on classification and regression tasks and whose adaptation to ranking problems leads to sub-optimal solutions. In turn, approaches that expressly consider ranking as the output out their computation have been shown to provide better results in both Information Retrieval (Joachims, 2006) and Recommender Systems (Shi et al., 2012a). From this point of view, recommendation algorithms targeting the ranking task can be seen as *Learning to Rank* methods, while approaches aiming to the rating prediction problem can be seen as solving a regression task and, therefore, they are expected not to perform as good as the *Learning to Rank* alternatives.

Offline evaluations, despite providing an affordable and accessible means to assess the quality of recommendations, have a limited usefulness since they cannot provide direct information about many of the aspects involved in the satisfaction of the user with respect to recommendations, such as relevance, surprisal, engagement, etc. Further, offline evaluation assumes that past user behavior can model future behavior, thus ignoring many variables such as the shifts in user's interests or, even, the perception of the users for the recommendations caused by new recommendation algorithms. For this reasons, offline evaluations need to be complemented with online evaluations, in which the performance of a new system is evaluated with real user feedback.

User Studies

One direct way of evaluating the performance of a recommender system consists in performing **user studies**, in which a set of test subjects is asked to interact with the recommendations provided by the tested system. Many works include the use of user studies in the evaluation of Recommender Systems. For example, Pu et al. (2011) and Knijnenburg et al. (2012) proposed different evaluation frameworks for the evaluation with user studies. Ziegler et al. (2005) and Ekstrand et al. (2014) conducted user studies to evaluate user perception with respect to diversity and novelty in recommendations. In user studies, while the user interacts with the system, a number of qualitative measurements can be gathered and, additionally, qualitative questions in the form of surveys can be carried out throughout the experiment. Clearly, user studies provide much more information than offline evaluations for each user. Nonetheless, user studies have also several limitations that may hinder their applicability to the evaluation of Recommender Systems. First, user studies are very expensive to conduct: recruiting a sufficiently large base of users is not a trivial task, and frequently involves monetary rewarding mechanisms. Second, participants need to adequately represent the population of users of the real system, thus covering different population strata in terms of gender, age, education, expertise in the recommendation domain, etc. Lastly, user studies must take into account that the results will be inherently biased as the users are aware that they are participating in an experiment.

Online Experiments

More indirectly, many business and organizations run controlled experiments on their systems in the form of A/B testing or alternative techniques (Kohavi et al., 2009). Usually, these experiments redirect a fraction of the traffic of a platform towards the evaluated system and measure system performance by means of user engagement metrics such as page views, click-through rate (Garcin et al., 2014) or, more directly, the economic benefit of the system (Shani et al., 2005). This kind of evaluation provides the strongest evidence as it is performed in real settings with real users. However, results of this kind of experiments must be analyzed carefully to draw reliable conclusions and discard differences of the evaluated caused by external factors or chance. Moreover, there is a risk involved in performing evaluations in real systems, as tested under-performing systems may affect negatively the experience of real customers.

2.2.3.2 *User vs. Business-Oriented Evaluation*

Another dimension of the evaluation of Recommender Systems concerns the side of the recommendation system that is being assessed. On one hand, the users need to be satisfied with the recommendations they receive. On the other hand, recommender systems are typically deployed by businesses or other organizations to increase revenue of the services they offer. In some sense, the user is not the end customer of a recommendation system, but is the business or organization that deploys it (Azaria et al., 2013).

While it is clear that providing a bad experience to the user affects negatively the performance of the system in terms of a business-oriented evaluation, the opposite may not be completely true. For instance, Netflix is known to avoid recommending new releases, however relevant to the users, since they have high costs to them (Shih et al., 2007). For that purpose, a complete evaluation of a recommender system has to consider not only the satisfaction of the users for the recommendations provided by the system, but also metrics that quantify the utility of the recommendations to the business or platform behind it.

2.2.3.3 *Accuracy and Alternative Quality Dimensions*

The evaluation of Recommender Systems can also be characterized by means of the property of interest that is sought to be enhanced. Traditionally, the evaluation of recommendation algorithms has been oriented towards maximizing the accuracy of recommendations, understood as retrieving as many relevant items as possible. Along with the progress targeting accuracy in Recommender Systems, researchers have realized that improving recommendations' usefulness and user satisfaction may require more than being accurate. In particular, Herlocker et al. (2004) stated that accuracy alone may not give users of recommender systems an effective and satisfying experience. McNee et al. (2006) and Ge et al. (2010) further specified that there are properties other than accuracy that have a larger effect on user satisfaction and performance, namely coverage, diversity, novelty, serendipity and assessment of user needs. Modeling and assessing user satisfaction is a complex task which involves several disciplines, such as Information Retrieval, User Modeling, Human-Computer Interaction, Marketing, Economics and Psychology.

In order to assess user satisfaction, new ways of evaluating Recommender Systems are being proposed. Bollen et al. (2010) analyzed the so-called "choice overload" i.e. the effect of presenting the user a large list of highly relevant items. In a user study they show that users, when asked to choose from a list with many relevant items, experience a difficulty in making a selection, which translates to a poor satisfaction with their final choice. The personality of users is another aspect to consider. As shown by Chen et al. (2013), personality can potentially quantify the need for diversity in the recommendations. Mood has also been found to be a major factor on how users interact with recommendations. Winoto and Tang (2010) showed that mood-aware recommendations can perform better in movie domains. In general, context, defined as all the possible variables that may influence the user preference for a given item in a certain situation, plays a crucial role in generating appropriate recommendations (Adomavicius and Tuzhilin, 2010). The capability of a Recommender System to explain a recommendation is also important. Herlocker et al. (2000) proposed a model for explaining recommendations based on the user's conceptual model of the recommendation process. Recommending novel items helps addressing the so-called *long tail effect* stated by Anderson (2006), caused when a few items are extremely popular and the rest are much less known. Finally, diversity in Recommender Systems, that is, addressing the user's varied tastes and her need for diverse recommendations, has been shown to help leveraging revenues from market niches Fleder and Hosanagar (2009), improve the attractiveness and usefulness of recommendations (Bollen et al., 2010; Pu et al., 2011) and avoid the so-called "filter bubbles" (Pariser, 2011).

2.2.4 *Issues, Limitations and Challenges*

The current state in the research and development of Recommender Systems has undoubtedly contributed to enhance user satisfaction and business success in varied scenarios. However, there are still many open issues, limitations and challenges that limit the usefulness of recommendations and are the object of active research in the topic. We review now several of these problems and how they affect the performance of recommendations.

Knowledge acquisition of user preferences is a crucial problem in Recommender Systems. Personalized recommendations require knowledge about the tastes of the user in the recommendation domain, typically in the form of preferences of the user for some items of the recommendation domain. Initially, when a user joins a recommender system, nothing or very little is known about what the user likes or is interested in. This is commonly known as the **cold start** problem (Kluver and Konstan, 2014). Moreover, collected knowledge about the interests of the user may be incomplete or biased and thus generate recommendations that do not satisfy user requirements. The task of gathering knowledge about the user, known as **active learning** (Elahi et al., 2014), is therefore determinant to provide successful recommendations.

Recommendation algorithms, specially content-based methods, are known to be **over-specialized** in the sense of recommending items that are *too* similar to what user already knows (Adamopoulos and Tuzhilin, 2014). Although similarity to previous user experiences is a good predictor of user relevance, it may provide very little added value to the real utility of recommendations, which is usually linked with their capacity allowing the discovery of new, unexpected content. In a similar manner, providing a list of very similar recommendations, however relevant, does also limit the quality of recommendations (Ziegler et al., 2005; Zhang and Hurley, 2008). In contrast, providing diverse recommendations is seen as a good strategy to address the variety of interest of the users and their need for varied recommendations.

Collaborative filtering algorithms are particularly known to suffer from a **popularity bias** towards recommending items in the long tail, an effect which is commonly known as the “Harry Potter Effect”[5]. There is a natural reason for this trend to begin with: collaborative filtering thrives on the populated regions of the user-item interaction matrix, and falls short in the sparser regions. Popular items live by definition in the more populated areas, since they carry more rating data that populates matrix cells, and collaborative algorithms are therefore more prone to end up recommending these items. The popularity bias of collaborative filtering algorithms has been studied by several authors. For instance, Zhao et al. (2013) show empirical evidence that popular items tend to be more recommended than not so popular ones, and proposes methods to alleviate this effect. Steck (2011) examined this issue in further depth and justified this popularity bias by the selection bias towards popular items in the available data.

Finally, when recommender systems are used at industrial scale, **scalability** becomes a crucial issue (Amatriain, 2012). Being able to provide recommendations in real time for millions of user poses challenges in terms of systems and architecture for data storage and algorithmic computation.

[5] http://recsyswiki.com/wiki/Harry_Potter_effect

2.2.5 *Software and Services*

The increasing interest in area the Recommender Systems has been accompanied by a rise in the available number of software and tools for the research, development and deployment of recommendation technologies. In this section, we review some of the software and services available for using recommendation technologies in both academic and professional domains.

One one hand, it is becoming more and more common that research groups release implementations of their recommendation algorithms to promote the diffusion of their work and, in general, to contribute to the progress in the field. Among these contributions, the release of open-source recommendation frameworks is specially relevant, since they provide a common infrastructure for the development of new algorithms under common structures and generic functionalities. This is the case, for example, of Easyrec[6], Lenskit (Ekstrand et al., 2011), Mahout (Owen et al., 2011) or MyMediaLite (Gantner et al., 2011). There has been noted though that the implementation of popular recommendation algorithms and, most particularly, evaluation methodologies may significantly vary from one framework to another, making the comparison and reproducibility of experiments in different frameworks a non-trivial task. For that purpose, Said and Bellogín (2014) presented RiVal, a recommender system evaluation toolkit that provides a complete control of the different evaluation dimensions of recommendation experiments: data splitting, evaluation strategies and computation of metrics.

On the other hand, there is an increasing number of companies providing recommendation services as a business model. This is the case of, among others, Gravity R&D[7], which provides recommendation technologies for a variety of recommendation scenarios, plista[8], which specializes in content recommendations for online publications, and YOOCHOSE[9], which specializes in providing recommendations for e-commerce, news and media.

One of the byproducts of this thesis is **RankSys**, a new framework for the implementation and evaluation of recommendation algorithms and techniques that specializes in the assessment and enhancement of novelty and diversity in Recommender Systems. To date, the part of this framework related to the assessment and enhancement of novelty and diversity in Recommender Systems has been released as open-source software[10], with plans to progressively release the rest of the framework, which includes efficient and parallelized implementations of many well known recommendation algorithms, as soon as the software is conveniently

[6] http://easyrec.org
[7] http://gravityrd.com
[8] http://plista.com
[9] http://yoochoose.com
[10] https://github.com/saulvargas/RankSys

documented. The high-level documentation of the current release of RankSys is available in Appendix A.

2.3 NOVELTY AND DIVERSITY IN RECOMMENDER SYSTEMS

Accurately predicting the users' interests was the main direct or implicit drive of the Recommender Systems field in roughly the first decade and a half of the field's development. A wider perspective towards recommendation utility, including but beyond prediction accuracy, started to appear in the literature by the beginning of the 2000's (Herlocker et al., 2004; Smyth and McClave, 2001), taking views that began to realize the importance of novelty and diversity, among other properties, in the added value of recommendation (McNee et al., 2006; Ziegler et al., 2005). This realization grew progressively, reaching an upswing of activity by the turn of the past decade (Adamopoulos and Tuzhilin, in press; Adomavicius and Kwon, 2012; Celma and Herrera, 2008; Hurley and Zhang, 2011). Today we might say that novelty and diversity are becoming an increasingly frequent part of evaluation practice. They are being included increasingly often among the reported effectiveness metrics of new recommendation approaches, and are explicitly targeted by algorithmic innovations time and again. And it seems difficult to conceive progress in the recommender systems field without considering these dimensions and further developing our understanding thereof. Even though dealing with novelty and diversity remains an active area of research and development, considerable progress has been achieved in these years in terms of the development of enhancement techniques, evaluation metrics, methodologies, and theory, and we deem the area is therefore ripe for a broad overview as we undertake in this section.

Novelty can be generally understood as the difference between present and past experience, whereas diversity relates to the internal differences within parts of an experience. The difference between the two concepts is subtle and close connections can in fact be established, depending on the point of view one may take, as we shall discuss. The general notions of novelty and diversity can be particularized in different ways. For instance, if a music streaming service recommends us a song we have never heard before, we would say this recommendation brings some novelty. Yet if the song is, say, a very canonical music type by some very well known singer, the involved novelty is considerably less than we would get if the author and style of the music were also original for us. We might also consider that the song is even more novel if, for instance, few of our friends know about it. On the other hand, a music recommendation is diverse if it includes songs of different styles rather than different songs of very similar styles, regardless of whether the songs are original or not for us.

The motivations for enhancing the novelty and diversity of recommendations are manifold, as are the different angles one may take when seeking these qualities. This is also the case in other fields outside information systems, where novelty and diversity are recurrent topics as well, and considerable efforts have been devoted to casting clear definitions, equivalences and distinctions. We therefore start by overviewing the reasons for and the possible meanings of novelty and diversity in Recommender Systems, with a brief glance at related perspectives in other disciplines. Then, we focus on the different perspectives and notions that we identify in the literature of novelty and diversity in Recommender Systems with an special interest in the proposals for the evaluation and promotion of novelty and diversity in recommendations.

2.3.1 *Why Novelty and Diversity in Recommendation*

Bringing novelty and diversity into play as target properties of the desired outcome means taking a wider perspective on the recommendation problem concerned with final actual recommendation utility, rather than a single quality side such as accuracy (McNee et al., 2006). Novelty and diversity are not the only dimensions of recommendation utility one should consider aside from accuracy (see Section 2.2.3.3), but they are fundamental ones. The motivations for enhancing novelty and diversity in recommendations are themselves diverse, and can be founded in the system, user and business perspectives.

From the system point of view, user actions as implicit evidence of user needs involve a great extent of uncertainty as to what the actual user preferences really are. User clicks and purchases are certainly driven by user interests, but identifying what exactly in an item attracted the user, and generalizing to other items, involves considerable ambiguity. On top of that, system observations are a very limited sample of user activity, whereby recommendation algorithms operate on significantly incomplete knowledge. Furthermore, user interests are complex, highly dynamic, context-dependent, heterogeneous and even contradictory. Predicting the user needs is therefore an inherently difficult task, unavoidably subject to a non-negligible error rate. Diversity can be a good strategy to cope with this uncertainty and optimize the chances that at least some item pleases the user, by widening the range of possible item types and characteristics at which recommendations aim, rather than bet for a too narrow and risky interpretation of user actions. For instance, a user who has rated the movie "Rango" with the highest value may like it because – in addition to more specific virtues – it is a cartoon, a western, or because it is a comedy. Given the uncertainty about which of the three characteristics may account for the user preference, recommending a movie of each genre generally pays off more than recommending, say three cartoons, as far as three hits do not necessarily bring three times the gain of one hit – e.g. the user might rent just one recommended movie anyway –, whereas the loss involved in zero hits is considerably worse than achieving a single hit. From this viewpoint we might say that diversity is not necessarily an opposing goal to accuracy, but in fact a strategy to optimize the gain drawn from accuracy in matching true user needs in an uncertain environment.

On the other hand, from the user perspective, novelty and diversity are generally desirable *per se*, as a direct source of user satisfaction. Consumer behaviorists have long studied the natural variety-seeking drive in human behavior (McAlister and Pessemier, 1982). Novel and diverse recommendations enrich the user experience over time, helping expand the user's horizon. It is in fact often the case that we approach a recommender system with the explicit intent of discovering something new, developing new interests, and learning. The potential problems of the lack of diversity which may result from too much personalization has recently come to the spotlight with the well-known debate on the so-called "filter bubble" (Pariser, 2011). This controversy adds to the motivation for reconciling personalization with a healthy degree of diversity.

Diversity and novelty also find motivation in the underlying businesses in which recommendation technologies are deployed. Customer satisfaction indirectly benefits the business in the form of increased activity, revenues, and customer loyalty. Beyond this, product diversification is a well-known strategy to mitigate risk and expand businesses (Lubatkin and Chatterjee, 1994). Moreover, selling in the long tail is a strategy to draw profit from market niches by selling less of more and getting higher profit margins on cheaper products (Anderson, 2006).

All the above general considerations can be of course superseded by particular characteristics of the specific domain, the situation, and the goal of the recommendations, for some of which novelty and diversity are indeed not always needed. For instance, getting a list of similar products (e.g. photo cameras) to one we are currently inspecting may help us refine our choice among a large set of very similar options. Recommendations can serve as a navigational aid in this type of situation. In other domains, it makes sense to consume the same or very similar items again and again, such as grocery shopping, clothes, etc. The added value of recommendation is probably more limited in such scenarios though, where other kinds of tools may solve our needs (catalog browsers, shopping list assistants, search engines, etc.), and even in these cases we may appreciate some degree of variation in the mix every now and then.

2.3.2 *Diversity in Other Fields*

Diversity is a recurrent theme in several fields, such as Sociology, Psychology, Economy, Ecology, Genetics or Telecommunications. One can establish connections and analogies from some – though not all – of them to Recommender Systems, and some equivalences in certain metrics, as we will discuss.

Diversity is a common keyword in Sociology referring to cultural, ethnic or demographic diversity (Levinson, 1998). Analogies to recommender system settings would apply to the user population, which is mainly a given to the system, and therefore not within our main focus here. In economy, diversity is extensively studied in relation to different issues such as the players in a market (diversity vs. oligopolies), the number of different industries in which a firm operates, the variety of products commercialized by a firm, or investment diversity as a means to mitigate the risk involved in the volatility of investment value (Lubatkin and Chatterjee, 1994). Of all such concepts, product and portfolio diversity most closely relate to recommendation, as mentioned in Section 2.3.1, as a general risk-mitigating principle and/or business growth strategy.

Behaviorist Psychology has also paid extensive attention to the human drive for novelty and diversity (McAlister and Pessemier, 1982). Such studies, especially the ones focusing on consumer behavior, provide formal support to the intuition that recommender systems users may prefer to find some degree of variety and surprise in the recommendations they receive, as discussed in Section 2.3.1.

An extensive strand or literature is devoted to diversity in Ecology as well, where researchers have worked to considerable depth on formalizing the problem, defining and comparing a wide array of diversity metrics, such as the number of species (richness), Gini-Simpson and related indices, or entropy (Patil and Taillie, 1982). Such developments connect to aggregate recommendation diversity perspectives that deal with sets of recommendations as a whole, as we shall discuss in Section 2.3.3.5.

Finally, the issue of diversity has also attracted a great deal of attention in the Information Retrieval field. A solid body of theory, metrics, evaluation methodologies and algorithms has been developed in this scope in the last decade (Agrawal et al., 2009; Carbonell and Goldstein, 1998; Chapelle et al., 2011; Chen and Karger, 2006; Clarke et al., 2008; Santos et al., 2010a; Zhai et al., 2003), including a dedicated search diversity task in four consecutive TREC editions starting in 2009 (Clarke et al., 2009, 2010, 2011b, 2012). Search and recommendation are different problems, but have much in common: both tasks are about ranking a set of items to maximize the satisfaction of a user need, which may or may not have been expressed explicitly. Consequently, we see a strong similarity between the problem of diversity in Recommender Systems and its analogous in Information Retrieval and wonder whether, as far as it were possible to draw models and principles from one area to the other, research on diversity in Recommender Systems might benefit from the insights and ongoing progress in search result diversification. In Section 2.4 we overview the work on search result diversification in Information Retrieval, upon which we develop an adaptation to Recommender Systems in Chapter 5.

2.3.3 *Perspectives on Novelty and Diversity*

Novelty and diversity are different though related notions, and one finds a rich variety of angles and perspectives on these concepts in the Recommender System literature. As pointed out at the beginning of this section, novelty generally refers, broadly, to the difference between present and past experience, whereas diversity relates to the internal differences within parts of an experience. Diversity generally applies to a set of items or "pieces", and has to do with how different the items or pieces are with respect to each other. Variants have been defined by considering different pieces and sets of items. In the basic case, diversity is assessed in the set of items recommended to each user separately (as in Ziegler et al. (2005)), and typically averaged over all users afterwards. But global diversity across sets of sets of items has also been considered, such as the recommendations delivered to all users (Adomavicius and Kwon, 2012, 2014; Zhou et al., 2010), recommendations by different systems to the same user (Bellogín et al., 2013), or recommendations to a user by the same system over time (Lathia et al., 2010). We now provide precise definitions of the perspectives on novelty and diversity in Recommender Systems that we have identified in the related work and review the most relevant work for their assessment and enhancement.

The following definitions cover a wide range of notions and perspectives on novelty and diversity involved in recommendations, but one might also study the diversity (in tastes, behavior, demographics, etc.) of the end-user population, or the product stock, the sellers, or in general the environment in which recommenders operate. While some works in the field have addressed the diversity in user behavior (Fleder and Hosanagar, 2009; Szlávik et al., 2011), we focus on those aspects a recommender system has a direct hold on, namely the properties of its own output.

2.3.3.1 *Long Tail Novelty*

The global, non-personalized perspective of Long Tail Novelty (Celma and Herrera, 2008; Park and Tuzhilin, 2008; Zhou et al., 2010) considers how novel are the items based on their popularity. As discussed in Section 2.2.4, recommendation algorithms may be biased towards recommending very popular items, the so-called short head. We argue that recommending popular items, however relevant, decreases the potential utility of Recommender Systems as tools for the discovery and exploration of vast catalogs since such popular items are typically found by means other than recommendations. Consider the case of movie recommendation where a system recommends to the user the blockbusters of the month: there is a high chance that the user is already aware of such movies, since such items are typically advertised and publicized in television, press and other mass media.

Long Tail Novelty, as defined, is concerned with providing less popular, obvious recommendations. Under this perspective, an item is novel if few people are aware it exist, i. e. the item is far in the long tail of the popularity distribution (Celma and Herrera, 2008; Park and Tuzhilin, 2008). Zhou et al. (2010) modeled popularity as the probability that a random user would know the item. To get a decreasing function of popularity, the negative logarithm provides a nice analogy with the inverse document frequency (IDF) in the vector-space Information Retrieval model, with users in place of documents and items instead of words, which has been referred to as inverse user frequency (IUF) (Breese et al., 1998). Based on the observed user-item interaction, (Zhou et al., 2010) proposed a metric that averages the inverse user frequency of the items in a recommendation R, which we call Mean Inverse User Frequency (MIUF):

$$\mathrm{MIUF}(R) = -\frac{1}{|R|}\sum_{i \in R} \log_2 \frac{|\mathcal{U}_i|}{|\mathcal{U}|} \tag{2.1}$$

where $\mathcal{U}_i$ is the set of users who know item i. Although the IUF formula has a reminiscence of the self-information measure of Information Theory, only for that to be properly the case, the probability should add to 1 over the set of items, which is not the case here.

Regarding the optimization of Long Tail Novelty in recommendations, Zhou et al. (2010) proposed algorithms that enhance the Long Tail Novelty of recommendations by means of hybrid strategies that combine collaborative filtering with graph spreading techniques. Lee and Lee (2013) used the concept of "experts and novices" to promote the novelty of recommendations. In a collaborative filtering setting, the authors do a clustering of the items and assign users to the cluster in which they have most of their ratings, that is, their clusters of expertise. Then, given a cluster for which a target user is novice (not expert), the knowledge of experts on that cluster is used to generate accurate yet novel recommendations. Ribeiro et al. (2012) applied evolutionary-inspired hybridization techniques to combine the outputs of different recommendation algorithms to maximize the Pareto-efficiency of accuracy, diversity and novelty.

Celma and Herrera (2008) took an interesting alternative view on Long Tail Novelty. Rather than assessing novelty just in terms of the long tail items that are directly recommended, they analyzed the paths leading from recommendations in the long tail through similarity links for collaborative filtering and content-based algorithms. For the case of collaborative filtering recommendations, the topology of the item similarity network leads to poor discovery ratio. On the other hand, content-based recommendations can provide more novel recommendations with lower perceived quality. Solutions suggested include promoting unknown artists from the long tail of the popularity distribution or selecting collaborative filtering or content-based recommendations depending on the users' needs.

2.3.3.2 *Unexpectedness*

A related but different notion considers the Unexpectedness (Murakami et al., 2008; Zhang et al., 2012; Adamopoulos and Tuzhilin, in press) involved in receiving recommendations that are novel in the sense that they are different or unfamiliar to the user experience. Adjectives such as unexpected, surprising and unfamiliar have been used to refer to this variant of novelty. Also, the notion of serendipity is similarly used to mean unexpectedness plus a positive emotional response – in other words, an item is serendipitous if it is novel and relevant. Unexpectedness differs from Long Tail Novelty, for which the novelty of an item is seen as independent of the target user, in considering the specific experience of a user when assessing the novelty carried by an item that is recommended to her, since the degree to which an item is more or less familiar can greatly vary from one user to the next.

Murakami et al. (2008) and Ge et al. (2010) considered that Unexpectedness is related to the difficulty of an item being predicted to a user. To estimate the difficulty of predictions, they compare the recommendations provided by a recommendation algorithm to those provided by a *primitive prediction method*. Examples of such *primitive methods* are predictions based on the viewing time-frame, favorite genres or celebrities of the users. The Unexpectedness is then measured by Ge et al. (2010) as the proportion of relevant items in a recommendation that cannot be obtained by a primitive method:

$$\mathrm{Unexp}_1(R) = \frac{|R \setminus PM|}{|R|}$$

where PM is set of items predicted by the primitive method. A refinement of the previous metrics considers the *serendipity* as the unexpectedness provided by the items that not only are unexpected but also relevant to the user:

$$\mathrm{Srdp}(R) = \frac{|(R \setminus PM) \cap Rel|}{|R|}$$

where Rel is the set of items that the user finds relevant.

Adamopoulos and Tuzhilin (in press) argued that these previous metrics do not fully capture Unexpectedness since these primitive methods do not necessarily take into account the expectations of the user. Rather than defining Unexpectedness with respect to some primitive prediction, they consider the set E_u of expected or obvious items the user would expect. The expected items for each user u can be defined in various ways, such as the set of items previously known by the user, items that are similar to those known by the user, or as a set of "typical" recommendations that she expects to receive or has received in the past. Based on this, they adapt first the previous metrics to consider the set E_u rather than the predictions of a primitive model:

$$\mathrm{Unexp}_2(R_u) = \frac{|(R_u \setminus E_u) \cap \mathrm{Rel}_u|}{|R|}$$

Additionally, the authors proposed a relaxed variant in which the distance to the expected set E_u of the items in the recommendation is taken as the measure of Unexpectedness:

$$\mathrm{Unexp}_3(R_u) = \frac{1}{|R|} \sum_{i \in R} \mathrm{dist}(i, E_u) \tag{2.2}$$

where dist is a distance between items based on common features of the items.

Together with proposals for assessing Unexpectedness, there is a variety of methods for its enhancement. For instance, Onuma et al. (2009) considered the bipartite graph defined between users and items by their preferences, and recommend those items bridging different groups of users and items to promote surprisal. Zhang et al. (2012) introduced the *Auralist* recommendation framework, one of whose components aims at finding recommendations that lie on the edge of the clusters defined by the similarities between user's preferences. Adamopoulos and Tuzhilin (in press) proposed a method to select items with high predicted relevance and whose distance to the expected set E_u is within some defined limits by combining the scores of predicted relevance and unexpectedness.

2.3.3.3 *Temporal Novelty*

User perception of novelty can also be considered within the interactions of the user with a recommender system over time. In this case, we define as Temporal Novelty (Lathia et al., 2010) the ability of the recommender systems not to repeat itself by providing the same or similar recommendations over time. This perspective evaluates the capacity of a recommender system at incorporating new knowledge about the user and adapting the recommendations to it.

Given a recommendation R_u^t provided by user u at time t, Lathia et al. (2010) proposed to measure its novelty as the ratio of items that were not recommended before:

$$\mathrm{TN}(R_u^t) = \frac{\left|R_u^t \setminus \bigcup_{\tau<t} R_u^\tau\right|}{|R_u^t|} \tag{2.3}$$

The metric gives a perspective of the ability of a recommender system to evolve with the changes in the environment in which it operates, rather than presenting users the same set of items over and over again.

Lathia et al. (2010) carried out two experiments. One online experiment showed that the users' perception of the recommendations lists degrades if they do not show novelty with respect to past recommendations. Another offline experiment compared the temporal novelty of some collaborative filtering recommendation algorithms over time, reaching interesting conclusions:

- Item-based neighbors recommendations have on average higher temporal diversity than matrix factorization approaches.
- Users with large profiles receive less novel recommendations.
- The more a user interacts with the system in a session, the more novel the next recommendations will be.
- Even when a specific user does not interact with the system for a certain period of time, the interactions of other users will bring her more temporal novelty.

These observations provide evidence that improving the Temporal Novelty of collaborative filtering recommendations is necessary. The authors propose two methods for maximizing Temporal Novelty:

- Switching between recommendation algorithms, taking advantage of the differences between algorithms while maintaining a high degree of accuracy.
- Randomly reranking recommendation lists by replacing a specific amount of top-N recommendations with others of lower predicted preference but more diverse with respect to previous recommendations.

2.3.3.4 *Intra-List Diversity*

Intra-List Diversity (Smyth and McClave, 2001; Zhang and Hurley, 2008; Ziegler et al., 2005) considers how different are the items in a recommendation between each other. This perspective is one of the most studied in the literature, and is concerned with addressing the need of users for varied recommendations – specially by avoiding redundant or mono-thematic suggestions (Zhang and Hurley, 2008) –, covering the user's complete spectrum of interests (Ziegler et al., 2005) and minimizing the risk in the recommendation (Wang and Zhu, 2009).

Perhaps the most frequently considered Intra-List Diversity metric and the first to be proposed is the so-called Intra-List Distance (ILD) (Smyth and McClave, 2001;

Algorithm 2.1 Greedy re-ranking of Ziegler et al. (2005)

```
$S \leftarrow \emptyset$
while $|R \setminus S| > 0$ do
  $i^* \leftarrow \arg\max_{i \in R \setminus S} (1 - \lambda)\ s(u, i) + \lambda\ \min_{j \in R} \mathrm{dist}(i, j)$
  $S \leftarrow S \cup \{i^*\}$
end while
return $S$
```

Zhang and Hurley, 2008; Ziegler et al., 2005). This metric is defined as the average pairwise distance of the items in a recommendation set:

$$\mathrm{ILD}(R) = \frac{1}{|R|\,(|R|-1)} \sum_{i,j \in R_u} \mathrm{dist}(i, j) \tag{2.4}$$

The computation of ILD requires defining a distance measure $\mathrm{dist}(i, j)$, which is thus a configurable element of the metric. Given the profuse work on the development of similarity functions in the recommender systems field, it is common, handy and sensible to define the distance as the complement of well-understood similarity measures, but nothing prevents the consideration of other particular options. The distance between items is generally a function of item features (Ziegler et al., 2005), though the distance in terms of interaction patterns by users has also been considered sometimes (Ribeiro et al., 2012; Veloso et al., in press).

The ILD scheme in the context of recommendation was first suggested, as far as we are aware of, by Smyth and McClave (2001), and has been used in numerous subsequent works (Veloso et al., in press; Zhang and Hurley, 2008; Ziegler et al., 2005). Some authors have defined this dimension by its equivalent complement Intra-List Similarity (ILS) (Ziegler et al., 2005), which has the same relation to ILD as the distance function has to similarity, e.g. $\mathrm{ILD} = 1 - \mathrm{ILS}$ if $\mathrm{dist} = 1 - \mathrm{sim}$.

When measured by the similarities or distances between items in a recommendation, the optimization of Intra-List Diversity typically involves a trade-off between the relevance of the recommended items and their diversity. This can be formulated as follows:

$$R = \underset{S \subset \mathcal{I}, |S| = N}{\arg\max}\ (1 - \lambda) \sum_{i \in S} s(u, i) + \lambda \sum_{i \in S} \sum_{j \in S} \mathrm{dist}(i, j)$$

where N is the desired recommendation list size, λ is the parameter that controls the trade-off between relevance and diversity and $s(u, i)$ is the predicted relevance score of the baseline recommendation algorithm for user u and item i. Directly solving the previous formulation constitutes a NP-complete problem. To solve it, practical, efficient approximations have been proposed in the literature. For instance, Ziegler et al. (2005) proposed *topic diversification* (see Algorithm 2.1), a greedy selection algorithm that re-ranks the recommendations provided by a baseline recommendation algorithm. This algorithm, which is structurally equivalent to the Maximal Marginal Relevance (MMR) of Carbonell and Goldstein (1998), selects at every step the item from the original recommendation list that maximizes a linear combination of the relevance score and the minimum distance to the items already selected. Di Noia et al. (2014) elaborated on this greedy selection scheme to make an adaptive control the trade-off between relevance and diversity by analyzing the propensity towards diversity of each user. Alternatively, Zhang and Hurley (2008) posed the diversity optimization task as a quadratic optimization problem that substitutes the selection of a subset by a real-valued vector y^* that maximizes the relaxed real-valued selection of the relevance-diversity trade-off:

$$y^* = \underset{y \in R^{|\mathcal{I}|}, \|y\|^2 = N}{\arg\max} \; (1-\lambda) \sum_{i \in \mathcal{I}} y_i \; s(u,i) + \lambda \sum_{i \in \mathcal{I}} \sum_{j \in \mathcal{I}} y_i \; y_j \; \mathtt{dist}(i,j)$$

The same authors proposed in (Zhang and Hurley, 2009) the partition of user profiles for maximizing Intra-List Diversity. Their procedure goes as follows: a partition of the user profile is done so that the resulting clusters minimize the intra-cluster distance; then different recommendations are generated for each cluster; finally, the different recommendations are combined into a single, varied diversification. Hurley (2013) also explored the maximization of relevance and diversity in matrix factorization approaches. The previously commented evolutionary-inspired hybridization of Ribeiro et al. (2012) also considered diversity as one of the components of their multi-objective recommendation algorithm.

Item distance approaches are found in most evaluations of this perspective of diversity in Recommender Systems. However, alternative formulations can be found. For instance, Küçüktunç et al. (2013) proposed a graph-based metric in which the diversity of a set of recommended items is determined by the l-step expansion relevance ($\mathtt{exprel}_l$), that is, the sum of the relevance scores of the items that are up to distance l from the set of recommended items in the graph of items. For the optimization of this perspective, the authors proposed a greedy approach that maximizes at every step the coverage of items in the graph.

The Intra-List Diversity can also be seen as a way to minimize the risk involved in a recommendation. Kabutoya et al. (2013) took a "less is more" approach inspired by the work of Chen and Karger (2006) and defined a probabilistic model to diversify recommendation lists. The goal of their model is to minimize the risk in a recommendation list so that a user selects at least one of the recommended items. For that, given a ranked list, they pick items assuming that the previously ranked items are not relevant. Wang and Zhu (2009) used Modern Portfolio Theory to both maximize the expected relevance of the items in a recommendation and minimize the risk, measured as the variance of the overall effectiveness, that a ranking may have. Shi et al. (2012b) elaborated on the previous approach and used latent factors as a better source to estimate the variance of each user.

Subtopic retrieval metrics adapted from search result diversification have been used to evaluate Intra-List Diversity in recommendations (Shi et al., 2012b; Küçüktunç et al., 2013; Kabutoya et al., 2013; Belém et al., 2013; Su et al., 2013). In Chapter 5 we propose a formal and principled adaptation of such metrics, specially the so-called Intent-Aware metrics, to the recommendation setting. Additionally, we explore in Chapter 6 new formulations to measure Intra-List Diversity in domains where genres are used to measure diversity.

2.3.3.5 *Sales Diversity*

Sales Diversity (Adomavicius and Kwon, 2012, 2014; Bellogín et al., 2013) is concerned with "making the most of the catalog", that is, procuring that all or most products in the business catalog get purchased to some extent, rather than having sales concentrating around a few items. This concept has been formulated by Anderson (2006) as "selling less of more" as a shift from selling a few "hits" or popular items in the short head towards a huge number of niches in the long tail. As opposed to the previous perspectives, this is a business-oriented perspective whose evaluation has to be made for the recommender system as a whole, rather than in a per user basis.

Adomavicius and Kwon (2012, 2014) proposed measuring the so-called Aggregate Diversity, defined as the total number of items that the system S recommends:

$$\mathit{Aggr\text{-}div}(S) = \left| \bigcup_{u \in \mathcal{U}} R_u^S \right| \tag{2.5}$$

Aggregate Diversity is a relevant measure to assess to what extent an item inventory is being exposed to users. The metric, or close variations thereof, have also been referred to as item coverage in other works (Bellogín et al., 2013; Ge et al., 2010; Herlocker et al., 1999, 2004). This concept can be also related to traditional diversity measures such as the Gini coefficient, the Gini-Simpson's index, or entropy (Patil and Taillie, 1982), which are commonly used to measure statistical dispersion in such fields as Ecology (biodiversity in ecosystems), Economics (wealth distribution inequality), or Sociology (e.g. educational attainment across the population). Mapped to recommendation diversity, such measures take into account not just whether items are recommended to someone, but to how many people and how even or unevenly distributed. To this extent they serve a similar purpose as aggregate diversity as measures of the concentration of recommendations over a few vs. many items. For instance, Fleder and Hosanagar (2009) measure sales concentration by the Gini index, which Shani and Gunawardana (2011) formulate as:

$$\mathrm{Gini}(S) = \frac{1}{|\mathcal{I}|-1} \sum_{k=1}^{|\mathcal{I}|} (2k - |\mathcal{I}| - 1)\; p(i_k \,|\, S) \tag{2.6}$$

where $p(i_k \,|\, S)$ is the probability of the k-th least recommended item being drawn from the recommendation lists generated by a system S:

$$p(i \,|\, S) = \frac{|\{u \in \mathcal{U} \,:\, i \in R_u\}|}{\sum_{u \in \mathcal{U}} |R_u|} \tag{2.7}$$

The Gini index and aggregate diversity have been used in subsequent work such as (Jannach et al., 2013). Other authors (e.g. Szlávik et al. (2011) or Shani and Gunawardana (2011)) suggest the Shannon entropy with similar purposes:

$$H(S) = \sum_{i \in \mathcal{I}} p(i \,|\, S)\; \log_2 p(i \,|\, S) \tag{2.8}$$

Related to this, Zhou et al. (2010) observe the diversity of the recommendations across users. They define the Inter-User Diversity (IUD) metric as the average pairwise ratio of different items between recommendations to users:

$$\mathrm{IUD}(S) = \frac{1}{|\mathcal{U}|\,(|\mathcal{U}|-1)} \sum_{u,v \in \mathcal{U}} \frac{|R_u \setminus R_v|}{|R_u|} \tag{2.9}$$

Enhancement techniques for Sales Diversity include the re-ranking of recommendation lists generated by a baseline recommendation algorithm (Adomavicius and Kwon, 2012) by means of different criteria that correlate with Sales Diversity, such as Long Tail Novelty. The same authors also suggested in (Adomavicius and Kwon, 2011) the use of graph-based approaches. Similarly, Zhou et al. (2010) also proposed enhancing Sales Diversity by means of hybrid strategies that combine collaborative filtering with graph spreading techniques. Niemann and Wolpers (2013) presented a new collaborative filtering approach based on items' usage contexts that improves the aggregate diversity of the results. Finally, Wu et al. (2014) introduced the Diversity aware Personalized PageRank method to reduce the impact of resource popularity on recommendations and then generate more diverse and novel recommendations to users.

2.3.3.6 *Sales Novelty*

Last, we consider **Sales Novelty** (Bellogín et al., 2013) as a perspective on its own, that measures the ability of a recommender system to provide original and unique recommendations not offered by alternative systems. This perspective becomes useful when considering a platform or business that seeks to distinguish itself from the competition, or when selecting an algorithm to add to an ensemble.

With a similar metric structure to Inter-User Diversity, Bellogín et al. (2013) define the Inter-System Diversity (ISD) metric in terms of how different the output of a system is with respect to other systems, in settings where several recommenders are operating. This can be defined as the ratio of systems that do not recommend each item:

$$\mathrm{ISD}(R^S) = \frac{1}{|\mathcal{S}|} \sum_{\Sigma \in \mathcal{S}} \frac{\left|R^S \setminus R^\Sigma\right|}{|R^S|} \tag{2.10}$$

where $\mathcal{S}$ is the set of recommender systems in consideration, and R^Σ denotes the recommendation to the target user by a system $\Sigma \in \mathcal{S}$.

2.4 SEARCH RESULT DIVERSIFICATION

Research and development in Information Retrieval (Baeza-Yates and Ribeiro-Neto, 2011) has been traditionally focused on accuracy and relevance as targets for satisfying the user information need. However, as in Recommender Systems, there is an increasing concern for the need of something more than accuracy to maximize the practical utility and the effective value of the retrieved information. In particular, the diversity of search result lists has been recognized (Clarke et al., 2008) as an important ingredient to cope with query ambiguity and underspecification.

Quite often a typical short textual query can represent more than one concept or interpretation (the case is clear, for example, with acronyms or polysemic words), in which case the query is called ambiguous. Consider the query "apple", which could refer to the fruit, the computer industry corporation, a record label, and other less common **interpretations**. Users interested in one interpretation would not usually be interested in the others. Even when the query does identify a unique concept or entity, it may still be underspecified in the sense that it may have different aspects. Consider a query like "Mallorca", which clearly refers to an island in the Mediterranean Sea, but still involves uncertainty about the actual specific user interest behind the query, which might relate to general information about the island, touristic deals, the local football team, etc. In this case these different **aspects** or **facets** do not need to be mutually exclusive, that is, users may be interested in two or more of them. In this work we will refer to both interpretations and facets as ***subtopics***, since we shall deal with both in the same way – as generally do prior approaches in the state of the art literature.

Traditionally, Information Retrieval research has been built upon the Probability Ranking Principle (PRP), which states that "if an IR's system response to each query is a ranking of documents in order of decreasing probability of relevance, the overall effectiveness of the system will be maximized" (Robertson, 1997). While this principle has been of great utility in the research and development in Information Retrieval systems for decades, it does not take into account the diversity of search results. This issue has been identified by many authors such as Chen and Karger (2006), Clarke et al. (2008) or Zhai et al. (2003).

As a strategy to cope with ambiguity and underspecification, **search result diversification** techniques (Agrawal et al., 2009; Carbonell and Goldstein, 1998; Chapelle et al., 2011; Chen and Karger, 2006; Clarke et al., 2008; Santos et al., 2010a; Zhai et al., 2003) have been proposed to cover as many subtopics as possible while still retaining sufficient relevance to satisfy the user need. In this case, the traditional assumption of independent relevance of documents of the Probability Ranking Principle does not hold. Here, the quality of the retrieval system cannot be quantified as an aggregation of relevance of each retrieved document, but as a property of the whole set of retrieved documents. Therefore, the evaluation and enhancement of the diversity of search result lists requires the definition of new models and perspectives beyond the Probability Ranking Principle.

Current offline evaluation practice in Information Retrieval relies on standard test collections for experimentation, such as those provided in the context of the TREC Web tracks (Clarke et al., 2009, 2010, 2011b, 2012). The problem of search result diversification was acknowledged in the 2009 to 2012 TREC Web tracks in the form of a diversity task. In each diversity task, a total number of 50 different search topics were presented together with relevance judgments provided by human assessors for documents in the search collection. Each search topic was represented with a short query and a representative set of subtopics (interpretations or aspects) related to different user needs. Topics were categorized as ambiguous or faceted, depending on whether its subtopics refer to interpretations or aspects of the query. Figure 2.1 shows examples of such ambiguous or faceted queries.

It is important to stress that, as well as relevance judgments, the subtopics of each query are not known by the TREC competition participants – or by systems being tested in research experiments using these datasets – when retrieving and ranking documents, and they are only used for evaluating the systems' output. This means that systems targeting diversity of search result lists need to use alternative sources to promote the diversity of documents in a search result.

In the remaining of the section, we review the main contributions regarding metrics and diversification techniques that have been proposed to assess and enhance the diversity of search results.

```
<topic number="19" type="ambiguous">
  <query>the current</query>
  <description>
    I’m looking for the homepage of The Current, a program on Minnesota
    Public Radio.
  </description>
  <subtopic number="1" type="nav">
    Take me to the homepage of The Current, a program on Minnesota
    Public Radio.
  </subtopic>
  <subtopic number="2" type="nav">
    I’m looking for the homepage of The Current newspaper in New Jersey.
  </subtopic>
  <subtopic number="3" type="nav">
    I want to find the homepage of The Current newspaper in Hartford.
  </subtopic>
  <subtopic number="4" type="nav">
    I want to find the homepage of The Current magazine in San Antonio.
  </subtopic>
</topic>

<topic number="21" type="faceted">
  <query>volvo</query>
  <description>
    I’m looking for information on Volvo cars and trucks.
  </description>
  <subtopic number="1" type="nav">
    I’m looking for Volvo’s homepage.
  </subtopic>
  <subtopic number="2" type="inf">
    Find reviews of the Volvo XC90 SUV.
  </subtopic>
  <subtopic number="3" type="inf">
    Where can I find Volvo semi trucks for sale (new or used)?
  </subtopic>
  <subtopic number="4" type="inf">
    Find a Volvo dealer.
  </subtopic>
  <subtopic number="5" type="inf">
    Find an online source for Volvo parts.
  </subtopic>
</topic>
```

Figure 2.1: Examples of ambiguous and faceted queries from TREC 2009 Web track topics.

2.4.1 *Metrics*

A variety of metrics specifically defined for the assessment of search result diversity by means of subtopics has been proposed in the literature of Information Retrieval the last decade. Here we introduce the most commonly used metrics in the context of the evaluation of the diversity task in the TREC Web track.

2.4.1.1 *Subtopic Retrieval Metrics*

Zhai et al. (2003) presented a seminal study describing evaluation metrics, methods and experimental results concerning the *subtopic retrieval* problem. The first proposed metric is called Subtopic Recall (S-recall). This metric computes, for a search result list R_q for query q, the retrieved proportion of the possible subtopics of the query:

$$\text{S-recall}(R_q) = \frac{\left|\bigcup_{d \in R_q} \mathrm{subtopics}_d\right|}{n_{\mathrm{subtopics}}} \tag{2.11}$$

where $\mathrm{subtopics}_d$ is the set of subtopics covered by document d and $n_{\mathrm{subtopics}}$ the number of possible subtopics of query q. As S-recall may not be an easy-to-compare metric across topics – consider the fact that the number of subtopics and how they are covered by related documents is highly different depending on each topic –, the authors provide another metric called Subtopic Precision (S-precision) in order to account for the "intrinsic difficulty" of each topic. S-precision is defined for a given S-recall level r as:

$$\text{S-precision@r}(R_q) = \frac{\mathrm{minRank}(R_q^*, r)}{\mathrm{minRank}(R_q, r)}$$

where $\mathrm{minRank}(R_q, r)$ is the minimum rank with recall level r and R_q^* is an optimal system for $\mathrm{minRank}(\cdot, r)$.

In Chen and Karger (2006) S-recall is found to be a derivation of their k-call family of metrics. In fact, since S-recall is defined as the total relative amount of subtopics retrieved, it is equivalent to the average of 1-call metrics marginalized to each subtopic:

$$\text{S-recall}(R_q) = \frac{1}{n_{\mathrm{subtopics}}} \sum_{s \in \mathrm{subtopics}_q} 1\text{-call}(R_q \mid s)$$

where $1\text{-call}(R_q \mid s)$ is a metric that evaluates to 1 when at least one document in R_q covers the subtopic s, and 0 otherwise.

2.4.1.2 *Intent-Aware Metrics*

Agrawal et al. (2009) proposed a generalization of some standard Information Retrieval metrics to acknowledge the possible *intents* (analogous to subtopics) of a query. They do it by evaluating the relevance of a search result list for each subtopic separately and then combine these partial results into a single one. Hence, given a generic metric M – such as nDCG, MRR, MAP, ERR –, its intent-aware version M-IA is defined as:

$$\text{M-IA}(R_q) = \sum_s p(s \mid q)\, M(R_q \mid s)$$

where $p(s \mid q)$ is the probability that the subtopic s is the intended interpretation or facet behind the query q, and $M(R_q \mid s)$ is the marginalization of the original metric that considers relevant only those documents covering subtopic s. For example, the intent-aware version of the Expected Reciprocal Rank (ERR-IA) metric of Chapelle et al. (2009), which is one of the most frequently used intent-aware metrics, gets the following expression:

$$\text{ERR-IA}(R_q) = \sum_s p(s \mid q) \sum_{k=1}^{|R_q|} \frac{1}{k}\, p(rel \mid d_k, s) \prod_{j=1}^{k-1} \big(1 - p(rel \mid d_j, s)\big) \tag{2.12}$$

where d_k is the document ranked at position k in R_q and $p(rel \mid d_k, s)$ is the probability of relevance of document d_k with respect to subtopic s.

2.4.1.3 *Redundancy Penalization Metrics*

Clarke et al. (2008) stressed the fact that most Information Retrieval evaluation metrics, such as MAP or nDCG, assume that the relevance of each document can be judged in isolation, independently from other documents, thus ignoring important factors such as redundancy between documents and the uncertainty – in the sense of incompleteness or ambiguity – in the query. The design of evaluation metrics should be consequently coherent with the actual user requirements. For this purpose, the authors present a framework for assessing diversity and novelty based on cumulative gain. Under their point of view, the relevance gain $G(d_k)$ of the k-th document d_k for a user need should be considered in the light of documents ranked above position k. The authors assume that subtopics may occur independently for every document and query, and the assessment of positive relevance judgments of a document for a subtopic involves an uncertainty that can be modeled with a fixed probability α of success in the judgment. These assumptions result in the following formulation of the gain $G(d_k)$:

$$G(d_k) = \sum_s \mathrm{rel}(d_k \,|\, s) \prod_{j=1}^{k-1} \big(1 - \alpha \, \mathrm{rel}(d_j \,|\, s)\big)$$

where $\mathrm{rel}(d_k \,|\, s)$ is the binary relevance judgment of document d_k with respect to subtopic s. By plugging this redundancy-aware gain in nDCG, we get the metric known as α-nDCG:

$$\alpha\text{-nDCG}(R_q) = \frac{1}{\alpha\text{-iDCG}} \sum_{k=1}^{|R_q|} \frac{1}{\log_2(k+1)} \sum_s \mathrm{rel}(d_k \,|\, s) \prod_{j=1}^{k-1} \big(1 - \alpha \, \mathrm{rel}(d_j \,|\, s)\big) \tag{2.13}$$

where α-iDCG is the normalization factor set as the maximum possible value of α-nDCG for an ideal ranking.

Clarke et al. (2011a) proposed later on the unification of α-nDCG and others diversity metrics. Diversity is accommodated through a linear combination of measures computed on individual subtopics (see the description of Intent-Aware metrics of the previous section). Novelty is accommodated by penalizing redundancy. In fact, some of the already presented metrics can be explained under this unification. After conducting some experiments with the test collection of the TREC 2009 Web track, results indicate that these metrics work as intended. Concurrently, Chapelle et al. (2011) determined that α-nDCG is roughly equivalent to ERR-IA when one consider a uniform distribution of subtopics in $p(s \,|\, q)$ a logarithmic instead of reciprocal discount, and a probability of relevance $p(\mathrm{rel} \,|\, d, s) = \alpha \, \mathrm{rel}(d \,|\, s)$.

2.4.1.4 *Cumulative Proportionality*

Dang and Croft (2012) proposed to consider the proportion of covering documents for each subtopic in a recommendation list. They emphasized the need for covering each subtopic of the search query by offering a number of relevant documents proportional to the interest of the subtopic they cover. The basis for measuring this proportionality is the so-called disproportionality metric, defined as:

$$DP(R_q) = \sum_s \mathbb{1}_{v_s \geqslant k_s^{R_q}} (v_s - k_s^{R_q})^2 + \frac{1}{2} n_{NR}^2$$

where v_s is the expected number of documents that cover the subtopic s, k_s^R the actual number of documents, and n_{NR} the number of non-relevant documents. The authors propose, on top on DP, a Cumulative Proportionality metric (CPR) that is the basis of their study and has the following formulation:

Algorithm 2.2 Greedy re-ranking of search result lists.

```
$S \leftarrow \emptyset$
while $|R \setminus S| > 0$ do
    $d^* \leftarrow \arg\max_{d \in R \setminus S} f_{obj}(d \,|\, S)$
    $S \leftarrow S \cup \{d^*\}$
end while
return $S$
```

$$\mathrm{CPR}(R_q) = 1 - \frac{1}{|R_q|} \sum_{k=1}^{|R_q|} \frac{DP_k(R_q)}{iDP_k}$$

where DP_k is the disproportionality at cut-off k and iDP_k the maximum possible value of DP_k.

2.4.2 *Diversification Methods*

In the last section, evaluation metrics have been introduced to measure the effectiveness of retrieval systems in the search result diversification task. One common characteristic of these metrics, as opposed to traditional relevance metrics, is that the usefulness of a set of retrieved documents cannot be calculated anymore by the individual relevance of each document. The primary consequence of this is that there is no ranking principle akin to the Probability Ranking Principle for document relevance that provides uniform instruction on how to rank documents for diversity. Therefore alternative approaches must be applied to diversify search results. In particular, many of the proposed solutions rely on applying greedy re-ranking techniques (see Algorithm 2.2) to result lists provided by traditional retrieval systems. As we have already seen in Section 2.3.3.4 in a recommendation setting, such methods pick iteratively documents from an initial search result list according to an objective function $f_{obj}(d \,|\, S)$ that determines the diversity gain obtained when a candidate document d is added to the set of already re-ranked documents S.

A main difference between diversification techniques lies in the source of diversity used. As previously stressed, one of the characteristics of a subtopic-oriented evaluation of diversity in Information Retrieval is that the subtopics of a query are used solely for evaluation purposes, and therefore retrieval systems that seek to enhance the diversity of search result lists need to use other sources to determine diversity. According to the nature of this source of diversity, two different approaches to enhance the diversity of document lists have been established: those that rely on the comparison between documents to maximize the diversity between them, called *implicit* approaches, and those that rely in some external information to infer the subtopics behind an ambiguous or underspecified query, which are known as *explicit* approaches or, alternatively, **Intent-Aware** approaches as we shall know them in this thesis.

2.4.2.1 *Implicit Approaches*

Implicit diversification approaches aim at diversifying results lists by means of minimizing the similarity between documents of a result list, thus aiming to cover as many subtopics as possible.

One of the first diversification methods in Information Retrieval appears in (Carbonell and Goldstein, 1998), where a method for combining query relevance and the so called *information novelty* for text retrieval is presented. This method is appropriate for scenarios where information redundancy is often observed among relevant documents. The method, called Maximal Marginal Relevance (MMR), establishes greedy selection based on a trade-off between the relevance of a document for a given query and the amount of new information this document provides with respect to previously retrieved documents. The proposed greedy algorithm selects, at each rank level, the document d that maximizes the following expression:

$$f_{MMR}(d \mid S) = \lambda \, rel(d, q) - (1 - \lambda) \max_{d' \in S} sim(d, d')$$

where λ is a parameter taking values between 0 and 1, S the previously re-ranked documents, rel the relevance score of document d for query q and sim a similarity measurement between documents.

Using the parameter λ, one can tune the algorithm towards relevance or information novelty. In fact, relevance and information novelty are not always valued the same way for every scenario. While simple and intuitive, the idea of MMR of maintaining some value with respect to a query and being as different as possible to what has already been retrieved has been widely used in other publications in Information Retrieval and Recommender Systems (Ziegler et al., 2005).

Similar to the work of Carbonell and Goldstein (1998), Zhai et al. (2003) used language models with KL-divergence or simple mixture models to calculate document similarity.

2.4.2.2 *Explicit or Intent-Aware Approaches*

Explicit or Intent-Aware approaches operate on a different basis than that of their implicit counterparts. They attempt to identify the subtopics behind a query by means of, for example, a categorization or documents or query reformulations. Then, documents are selected as to maximize the coverage of the inferred subtopics while minimizing their redundancy. In the context of the TREC Web track, this family of methods, specially the xQuAD algorithm of Santos et al. (2010a), has proven to be more effective than their implicit counterparts.

Agrawal et al. (2009) assume that there is a taxonomy of information whose topical level models the possible subtopics of the queries, so documents and queries may belong to more than one category of the taxonomy. The authors also assume that usage statistics have been collected on the distribution of user intents over the categories. Using this knowledge, they develop an objective that trade-offs relevance and diversity to minimize the risk of dissatisfaction for the average user. Specifically, knowing the categories of the taxonomy both queries and documents belong, the usage statistics provide a way of determining the probability of a category belonging to a document, i.e., $p(c\,|\,q)$ and also the probability $V(d\,|\,c,q)$ of a document d satisfying the user intent represented by the category c the query q belongs to, they introduce the Intent-Aware Selection (IA-Select) algorithm to greedily select items from a initial search result list according to the following objective function:

$$f_{IA-Select}(d\,|\,S) = \sum_{c} p(c\,|\,q)\; V(d\,|\,c,q) \prod_{d' \in S} \big(1 - V(d'\,|\,c,q)\big)$$

where the set of documents S contains the documents previously selected by IA-Select in the previous steps.

In (Santos et al., 2010a) the Explicit Query Aspect Diversification (xQuAD) algorithm is presented. The xQuAD algorithm makes use of query reformulations provided by commercial web search engines to derive new sub-queries that will cover the possible aspects of the initial query. Given an ambiguous query q and a ranking of retrieved documents R_q, xQuAD greedily selects a new ranking S by maximizing at every step of the selection the following mixture probability:

$$f_{xQuAD}(d\,|\,S) = (1-\lambda)\; p(d\,|\,q) + \lambda\; p(d, \neg S\,|\,q)$$

where $p(d\,|\,q)$ is the probability of the document d being observed given the initial query q and $p(d, \neg S\,|\,q)$ the probability of observing the document d but not the documents already in S. Using the set of reformulations or sub-queries $\{q_r\}$ of query q, the authors develop $p(d, \neg S\,|\,q)$ by marginalizing it across sub-queries and assuming independence between documents given a sub-query, resulting in the following expression:

$$\begin{aligned}
p(d, \neg S\,|\,q) &= \sum_{q_r} p(q_r\,|\,q)\; p(d, \neg S\,|\,q_r) \\
&= \sum_{q_r} p(q_r\,|\,q)\; p(d\,|\,q_r)\; p(\neg S\,|\,q_r) \\
&= \sum_{q_r} p(q_r\,|\,q)\; p(d\,|\,q_r) \prod_{d' \in S} \big(1 - p(d'\,|\,q_r)\big)
\end{aligned}$$

In a later publication, Santos et al. (2010b) proposed a way to determine a way of selecting λ optimally for each query, adapting the specific need for diversification.

Analogously to the previous proposals, the proportionality framework of Dang and Croft (2012) introduces a greedy re-ranking strategy known as the Proportionality Method (PM), which is based on the system used to assign seats in legislative elections in some countries:

$$\begin{aligned} f_{PM}(i|S) = \lambda \, & \frac{v_s^*}{1 + 2\sum_{j \in S} \frac{p(j|s^*)}{\sum_{s'} p(j|s')}} \, p(i|s^*) \\ & + (1-\lambda) \sum_{s \neq s^*} \frac{v_s}{1 + 2\sum_{j \in S} \frac{p(j|s)}{\sum_{s'} p(j|s')}} \, p(i|s) \end{aligned} \tag{2.14}$$

where $s^* = \arg\max_s v_s / \left(1 + 2\sum_{j \in S} \frac{p(j|s)}{\sum_{s'} p(j|s')}\right)$ indicates the least-covered subtopic – in this case, replaced also by query reformulations – in S.

2.4.2.3 *Other Approaches*

The problem in diversity in Information Retrieval has also been approached from a *Learning to Rank* (Liu, 2009) point of view. The first reference found is (Radlinski et al., 2008), where two different algorithms, Ranked Explore and Commit and Ranked Bandits Algorithm, use data of user clicks to produce diverse rankings. Yue and Joachims (2008) also present a Learning to Rank approach for learning diverse subsets using structural SVM's. More recently, Slivkins et al. (2010) presented a scalable approach that takes into account document similarity and context with appropriate theoretical foundations.

Wang (2009) studied the problem of ranking under uncertainty using Modern Portfolio Theory. While the classic Probability Ranking Principle approaches deal with maximizing the effectiveness in ranked lists, they do not consider the implicit risk (measured as the variance of the overall effectiveness) that a given ranking may have. If the relevance of each document is considered a random variable, the expected value and the risk (variance) of the overall relevance of a ranked list may be jointly considered and optimized. In their paper, the authors show that this approach can improve the results for subtopic retrieval of standard and MMR-diversified approaches in terms of S-recall and other diversity metrics.

A diversity-aware alternative for PageRank (Brin and Page, 1998) called DivRank is presented in (Mei et al., 2010). As PageRank, DivRank is based on a random walk over a network of linked documents with a teleportation component and assumes that connected documents tend to be more similar than others whose linkage is weaker. The particularity of DivRank is that the transition probabilities from one document to another are adjusted at each step of the random walk to be proportional to the number of times the incoming document has been visited. This adjustment leads to a "rich gets richer" effect where nodes with a high probability absorb weaker neighbor nodes so when the iterations converge to a stationary state the documents with the highest probabilities.

3

EXPERIMENTAL DESIGN

3.1 INTRODUCTION

We introduce in this chapter a common design frame that is used as a general basis in the experiments of the following chapters. The experimental design is purposed to be as uniform and consistent as possible across the different parts of the work reported here, in order to facilitate the interpretation of results. This design furthermore aims to provide concise and clear guidelines to facilitate the reproducibility of our experiments and the comparison with our outcomes, avoiding particular design decisions which might provide unfair advantages to our proposed approaches.

The rest of the chapter is structured as follows. Section 3.2 presents the three collaborative filtering datasets on which we evaluate our proposals. In Section 3.3 we give details about the formulation of the baseline recommendation algorithms that we evaluate and enhance in terms of novelty and diversity. Then, Section 3.4 introduces the offline, rank-based evaluation methodology that we have followed in all our experiments. Finally, in Section 3.5 we show and comment some preliminary results of accuracy-based metrics on the described experimental design.

3.2 DATASETS

We have selected three different datasets for our experiments: MovieLens1M, Netflix and Million Song Dataset. These datasets cover two well-known recommendation domains: movie recommendation and music recommendation. Two of them are publicly available while the other was removed due to anonymity issues. One of them is small enough to facilitate the reproducibility of our experiments, while the others contain larger amounts of data to provide further proof of the generality (and scalability) of our proposals.

We provide details about these three datasets in this section. We focus on their main characteristics, namely the number of users and items and the type and amount of interactions between them, as well as other important properties such as their temporal distribution (when available) and their popularity biases. A summary of some of the reported magnitudes is shown in Table 3.1.

Additionally, since many of our experiments rely on some categorization of the items, we have used (movie and music) genres for each dataset. In the case of the

	$\lvert\mathcal{R}\rvert$	$\lvert\mathcal{U}\rvert$	$\lvert\mathcal{I}\rvert$	**density**	$\lvert\mathcal{G}\rvert$	$\lvert\mathcal{I}_{\mathcal{G}}\rvert$
ML1M	1,000,209	6,040	3,706	4.47%	18	3,661
Netflix	100,480,507	480,189	17,770	1.18%	28	9,320
MSD	48,590,563	1,129,318	379,962	0.01%	21	150,959

Table 3.1: Characteristics of the collaborative filtering datasets of our experimental design.

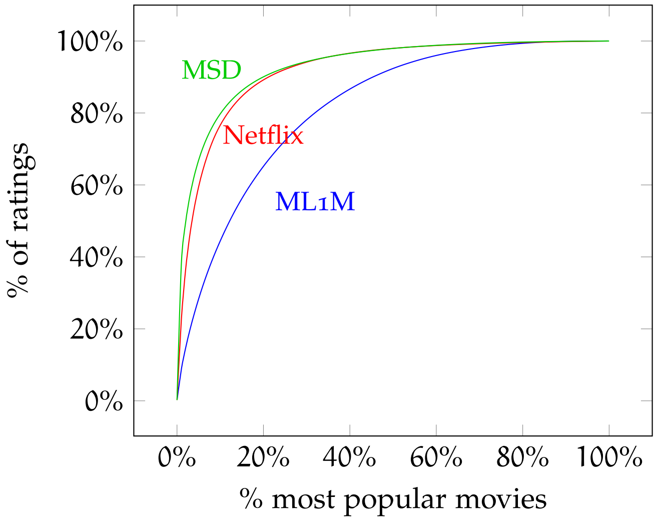

Figure 3.1: The popularity distribution of MovieLens1M, Netflix and Million Song Dataset.

MovieLens1M dataset, the genres of the items are already provided in the public dataset release, while in the others genres were extracted by us. We shall provide details of how we extracted such genres and some basic information about them.

3.2.1 *MovieLens1M*

The MovieLens1M dataset is the second-largest of the MovieLens datasets provided by GroupLens. It consists of 1 million ratings for 3,700 movies by 6,000 users. The ratings were made on a 1 to 5 stars scale by users who joined MovieLens in 2,000 and entered at least 20 ratings. The dataset includes information about the age and occupation of the users, and the title, the year of release and genres of the movies. There is a total of 3,661 movies with genre information covering a total of 18 different genres. Compared with the other datasets in our experimental design (see Table 3.1), it is a small but dense dataset.

Together with the value of the rating, the dataset also includes the time when each rating was entered. As Figure 3.2 shows, the ratings were collected over a period of three years, starting in April 2000 and ending in March 2003. However, most of the ratings were made in 2,000, indicating that most users stopped using

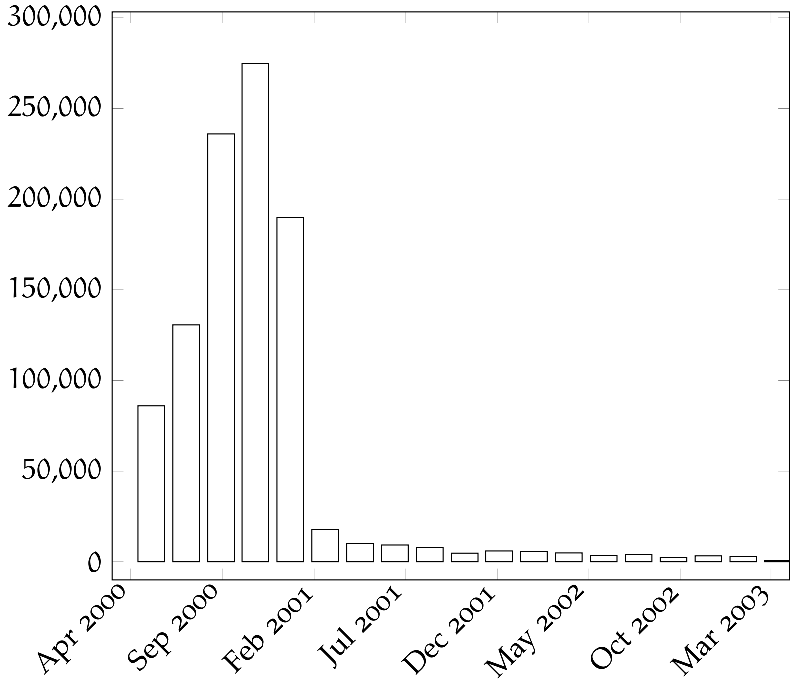

Figure 3.2: Temporal distribution of the ratings in MovieLens1M.

the service after a short period of time. This observation discourages the use of this dataset for time-based evaluations.

Together with the smaller MovieLens100k, this dataset is one of the most widely used for testing recommendation algorithms in the literature. It has the advantage of being sufficiently small so that computations can be performed in commodity hardware in a short time and still provide meaningful results due to its density and quality of data. This facilitates the reproducibility of our experiments and the comparison to results by other authors. For these reasons, we have selected it as the primary dataset in our experiments.

3.2.2 *Netflix*

To contrast and confirm the results for MovieLens in the movie recommendation scenario, we have included the dataset released for the Netflix Prize in our experiments. This dataset consists of 100 million ratings from over 480 thousand randomly-chosen users on nearly 18 thousand movie titles. The data were collected between October, 1998 and December, 2005 and reflect the distribution of all ratings received by Netflix during this period. The ratings are also on a scale from 1 to 5 stars.

In the case of the Netflix Prize, no genre information was provided in the data, so we extracted that information from IMDb. We were able to obtain the genres

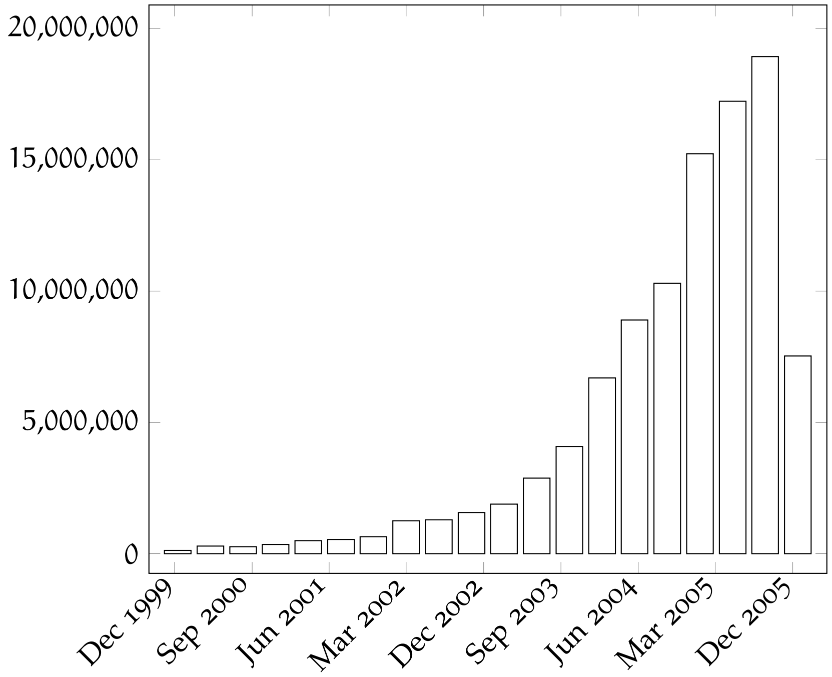

Figure 3.3: Temporal distribution of the ratings in Netflix.

of 9,320 movies resulting in a total of 28 different genres. The movies with genres account for almost 83% of the ratings.

The temporal distribution of the ratings of the Netflix dataset can be seen in Figure 3.3. In this case, the number of ratings increases over time, showing the increasing popularity of the platform. For instance, we can observe that most of the ratings are concentrated in the last two years. As done by Lathia et al. (2010), we consider this dataset adequate for a time-based splitting of the data for training and test purposes which we will apply when assessing the temporal novelty of recommendation algorithms in Chapter 4.

The popularity bias of the Netflix data is shown in Figure 3.1. The bias is in this case is higher than in MovieLens1M and similar to the Million Song Dataset. In particular, 20% of the most popular movies concentrate more than 80% of the total number of ratings.

3.2.3 *Million Song Dataset*

The most recent of the datasets we have tested is the Million Song Dataset. This dataset contains a collection of audio features and metadata for a million contemporary popular music tracks to promote the research in music Information Retrieval. The core of the dataset is the feature analysis and metadata for one million songs, provided by The Echo Nest. For our experiments, we only considered the Taste Profile subset, which includes 48 million playcount triplets by 1,100,000 users for 380,000 songs. As in the work of Aiolli (2013), we take binarized play counts since, as warned by the challenge organizers, play counts are unreliable and not necessarily correlate with likings. In this case, no temporal information is provided in the dataset. As Table 3.1 shows, this dataset is much sparser than the other two. This is expected as the number of items is much higher.

No genre information was provided for this dataset either. However, a complementary dataset provides tags extracted from Last.fm. We performed a simple hierarchical clustering to find the most general tags, which corresponded in a great proportion to potential music genres. Then, we manually selected those tags representing music genres and sub-genres and kept those that defined easily distinguishable styles and tastes while avoiding too general super-genres such as "Pop-Rock". The result consists of a set of 21 different genres covering almost 151,000 songs.

Finally, the popularity bias of this dataset, as Figure 3.1 shows, is slightly higher than that of the Netflix dataset.

3.3 RECOMMENDATION ALGORITHMS

We now provide details about the recommendation algorithms we have used across our experiments. We have focused on collaborative filtering algorithms as they have been shown to be quite effective in the datasets presented in the previous section. Although we have tested some content-based approaches, we have discarded them as they clearly underperform when compared to collaborative filtering alternatives.

As trivial baselines, we include random recommendations and a popularity-based recommender, which is the obvious baseline to beat in ranking tasks (Amatriain, 2013). These two non-personalized algorithms allow us to put in context the novelty and diversity of more elaborated recommendation algorithms: most-popular recommendations provide, by definition, the lowest Long-Tail Novelty and Sales Diversity while random recommendations are a natural source for novelty and diversity in all perspectives.

As personalized algorithms, we have used four different methods optimized for the ranking task that cover the main families in collaborative filtering: memory and model-based. Regarding the memory-based algorithms, we have considered both the user and item-based nearest neighbors algorithms. The specific selected variants, which we have found to be the most effective, in terms of precision-related metrics, along the years in carrying through all the work reported here, are based on the work of Cremonesi et al. (2010) and Aiolli (2013). In particular, our user-based (UB) algorithm computes its scores according to the following scoring function:

$$s_{UB}(u,i) = \sum_{v \in \mathcal{U}} \mathbb{1}_{v \in N_K(u)} \; sim(u,v) \; r_{u,i}$$

where $N_K(u)$ denotes the set of K most similar neighbors to user u. The similarity between users in based on the cosine similarity between the user profiles:

$$sim(u,v) = \frac{|\mathcal{I}_u \cap \mathcal{I}_v|}{\sqrt{|\mathcal{I}_u| \, |\mathcal{I}_v|}}$$

Analogously, the item-based (IB) variant used has the following formulation:

$$s_{IB}(u,i) = \sum_{j \in \mathcal{I}_u} \mathbb{1}_{i \in N_K(j)} \; sim(i,j) \; r_{u,j}$$

where $N_K(j)$ in this case denotes the set of K most similar neighbors to the item j in the user profile. The similarities between items are calculated using the cosine similarity between item profiles as well:

$$sim(i,j) = \frac{|\mathcal{U}_i \cap \mathcal{U}_j|}{\sqrt{|\mathcal{U}_i| \, |\mathcal{U}_j|}}$$

For the model-based family of recommendation algorithms, we have chosen two methods that represent two of the best-known variants in this family: matrix factorization and latent semantic analysis. In the first case, we used the implicit Matrix Factorization (iMF) of Hu et al. (2008), which factorizes the interaction matrix $\mathcal{R}$ into two matrices $P \in \mathbb{R}^{|\mathcal{U}|,k}$ and $Q \in \mathbb{R}^{|\mathcal{I}|,k}$ whose product determines the scoring function:

$$s_{MF}(u,i) = P_u \cdot Q_i^t$$

where P_u is the row vector of P that corresponds to user u and Q_i the row vector of Q which describes item i in the latent vector space. The matrices P and Q are obtained by an alternating minimization of the following weighted least squares loss function:

$$\mathcal{L}_{iMF}(P,Q) = \sum_{(u,i) \in \mathcal{U} \times \mathcal{I}} c_{u,i} \left(P_u \, Q_i^t - r_{u,i}\right)^2 + \lambda \left(\sum_u \|P_u\|^2 + \sum_i \|Q_i\|^2 \right)$$

where $r_{u,i} = 0$ when $(u,i) \notin \mathcal{R}$, λ is a regularization parameter and $c_{u,i}$ is the weight of the local error for the prediction of user u and item i. The weights $c_{u,i}$

	UB	IB	iMF			pLSA
	K	K	k	α	λ	k
ML1M	100	10	50	1	0.1	50
Netflix	100	10	50	10	0.1	50
MSD	200	20				

Table 3.2: Parameters chosen for the recommendation algorithms for each dataset.

are chosen to emphasize the importance of the observed interactions by using the following formula:

$$c_{u,i} = \begin{cases} 1 & (u,i) \notin \mathcal{R} \\ 1 + \alpha\, r(u,i) & (u,i) \in \mathcal{R} \end{cases}$$

Finally, we used the probabilistic Latent Semantic Analysis (pLSA) of Hofmann (2004). In this case, the scores are based on a joint probability $p(u,i)$ of observation of pairs of users and items. These probabilities are in turn obtained from a model $\theta = \{p(u|z), p(i|z), p(z)\}$ that considers a set $\{z\}$ of k latent variables that explain the interactions between users and items so that $p(u,i|z) = p(u|z)\; p(i|z)$. Therefore, by marginalizing each probability $p(u,i)$ by the set of latent variables we obtain the following scoring function:

$$s_{pLSA}(u,i) = p(u,i) = \sum_{z} p(u|z)\; p(i|z)\; p(z)$$

In this case, the model θ is obtained by means of an expectation-maximization algorithm that minimizes the following loss function:

$$\mathcal{L}_{pLSA}(\theta) = \sum_{(u,i)\in\mathcal{R}} r_{u,i}\; \log p(u,i)$$

Unless explicitly indicated, the specific parameters of the previous scoring functions for each dataset take the values shown in Table 3.2. Such values were manually chosen to optimize the precision of the recommendations.

As stated in Aiolli (2013), matrix factorization approaches are not effective in the Million Song Dataset, as our attempts at it confirmed, whereby we omit results with iMF in this dataset. For identical reasons, we also discard our results of the pLSA algorithm for the Million Song Dataset.

3.4 EVALUATION METHODOLOGY

Now we give details about how we performed the evaluation of the recommendations generated by the previous baseline recommendation algorithms and our proposals in the following chapters on the three datasets. For the reasons detailed in Section 2.2.3, we opt for a ranking-based evaluation. In such evaluation, two steps are involved: splitting the datasets into training and test subsets and selecting the items to rank.

For the MovieLens1M and Netflix datasets, we have performed a classic 5-fold cross-validation split, in which 4 folds are used for training and the remaining one for test. In the Million Song dataset we take the partition provided with the data release, which consists of test data for 110,000 of the users of the dataset.

In all cases, the recommenders to be evaluated are requested to produce recommendations (i.e. to rank items) for all users who have data in both the training and the test subsets. The items to rank for each user include all items having data in the training subset. Two restrictions are applied when selecting candidate items for these rankings: items in the training subset of the user are discarded, as are items without genre data. The latter restriction is motivated by our novelty and diversity-based evaluation. Some of the evaluated perspectives, namely Unexpectedness and Intra-List Diversity, require genre information for the items, therefore the restriction. For each ranking, we consider only the top-100 ranked items.

Finally, the resulting recommendation lists are evaluated in terms of ranking-oriented metrics. In particular, we consider as positive or relevant those items that appear in the test subset of the user and have a rating value above a particular threshold. In the case of MovieLens1M and Netflix, only the ratings with 4 and 5 are considered relevant and, in the Million Song Dataset, no threshold is applied so that every item in the user test is treated as relevant.

3.5 ACCURACY RESULTS

In Table 3.3 we show the evaluation results of the six baseline recommendation algorithms with the previously explained evaluation methodology in the three datasets. We evaluate precision (P) and the normalized Discount Cumulative Gain (nDCG) at cut-offs 20 and 50. Additionally, two quality measures are included: the average number of retrieved items (numRet) and the ratio of users for which the recommenders were able to provide recommendations (userCov). These two metrics provide a check for anomalous cases where recommenders are not able to provide enough or none recommendations for certain users.

In the MovieLens1M dataset, we see that random recommendations achieve the worst results in terms of accuracy. The second worst algorithm is the popularity-based, which clearly outperforms the random recommendations but falls considerably behind the personalized algorithms in terms of accuracy. Regarding the personalized algorithms, we observe that the iMF baseline gets the best results while the IB has the lowest performance according to all metrics and cutoffs. The pLSA and UB baselines offer comparable results that depend on the considered metrics: for the rank-unaware precision pLSA works better than UB, but in the graded, rank-aware nDCG, UB slightly outperforms pLSA.

		P@20	**P@50**	**nDCG@20**	**nDCG@50**	**numRet**	**userCov**
ML1M	**Rnd**	0.0057	0.0057	0.0052	0.0084	99.99	1.0000
	Pop	0.1215	0.0848	0.1561	0.1873	99.99	1.0000
	iMF	0.2335	0.1580	0.3394	0.3927	99.99	1.0000
	pLSA	0.2111	0.1454	0.2884	0.3403	99.99	1.0000
	UB	0.2055	0.1373	0.3039	0.3501	99.99	1.0000
	IB	0.1874	0.1272	0.2586	0.3035	99.52	1.0000
Netflix	**Rnd**	0.0022	0.0022	0.0018	0.0027	100.00	1.0000
	Pop	0.0909	0.0739	0.0960	0.1232	100.00	1.0000
	iMF	0.1778	0.1310	0.2345	0.2709	99.92	0.9992
	pLSA	0.1842	0.1305	0.2196	0.2486	99.92	0.9992
	UB	0.1923	0.1326	0.2425	0.2681	99.82	0.9992
	IB	0.1582	0.1170	0.1891	0.2209	94.38	0.9990
MSD	**Rnd**	0.0001	0.0001	0.0001	0.0001	100.00	1.0000
	Pop	0.0185	0.0128	0.0330	0.0421	100.00	1.0000
	UB	0.1018	0.0546	0.1898	0.2069	99.99	1.0000
	IB	0.1078	0.0602	0.1904	0.2119	78.25	0.9998

Table 3.3: Results of the baseline recommendation algorithms in terms of accuracy-based metrics, average recommended items and user coverage in MovieLens1M, Netflix and Million Song Dataset.

In the Netflix dataset, the results of the random and most-popular recommendations are comparable to those of MovieLens1M. Nevertheless, the results of the personalized algorithms differ. Overall, the UB algorithm has the best outcomes among the different metrics and cut-offs. The iMF algorithm has, however, better results than UB in nDCG@50, but is the second best in P@50 and nDCG@20. pLSA is the third best method, although it outperforms iMF in P@20. The IB method, as in MovieLens1M, has the worst performance.

Finally, the Million Song Dataset shows similar results for random and most-popular recommendations, but a large difference in personalized recommendations. In this case, the item-based algorithm works slightly better than the user-based, which contrasts with the results of the much denser movie recommendation datasets. The IB presents however the limitation of returning on average only about 78 recommended items for each user, failing to return at least 100 for each user as requested by the evaluation methodology.

4

A UNIFIED FRAMEWORK FOR NOVELTY AND DIVERSITY IN RECOMMENDER SYSTEMS

4.1 INTRODUCTION

In Chapter 2 we have presented a broad overview of the perspectives on novelty and diversity in Recommender Systems. In the last years an increasing stream of work in the field has resulted in a variety of proposals in the form of metrics and algorithms for taking into account these properties. While all this prior work has undoubtedly contributed to an increasing level of visibility and relevance of the topic, we miss a clear common methodological and conceptual ground for explaining and modeling novelty and diversity in recommendations. The absence of such common basis has in particular the following consequences:

- Unexplored connection between perspectives: as seen in Chapter 2, novelty or diversity do not refer to perfectly identifiable concepts, but they comprise a generality of definitions for different perspectives. These different perspectives are clearly different from one another, although some of them are naturally related. It would be desirable to explain the differences and connections between these perspectives with proper formal models.

- Lack of consensus on metrics: for a specific perspective on novelty or diversity in recommendations there may exist more than one possible metric for assessing it. Depending on the choice of a metric, different outcomes may result. Such outcomes may be equivalent in some cases, but they may also totally diverge. Identifying the equivalences and differences between metrics helps making better assessments on the performance of recommendation algorithms, and also contributes to the reproducibility of experiments.

- Lack of important properties such as position and relevance-awareness: considering the browsing models behind many of the relevance and diversity metrics in the field of Information Retrieval (Carterette, 2011; Clarke et al., 2008; Moffat and Zobel, 2008), it seems sensible to consider that novelty and diversity metrics for recommender systems should also take into account the position and relevance of the items composing the recommendations to determine the utility perceived by the user.

The present chapter aims to provide a formal ground for the unification of different perspectives to measure and enhance novelty and diversity. We propose a formal metric framework that unifies and generalizes several state of the art measures, and enhances them with configurable properties not present in previously reported evaluations. Specifically, the proposed scheme supports metrics that take into account the ranking and relevance of recommended items. These properties are introduced by taking into account how users interact with recommendations – top items get more attention – and user subjectivity – items the user does not like add little to the effective diversity of the recommendation, no matter how novel the items were objectively.

The proposed framework roots recommendation novelty and diversity metrics on a few ground concepts and formal models. We identify three essential concepts: discovery, relevance and choice, upon which the framework is built. The metric scheme takes at its core an item novelty model – discovery-based or distance-based – which mainly determines the nature of the resulting recommendation metric. Item rank and relevance are introduced through a probabilistic recommendation browsing model, building upon the same three basic concepts. Based on the combination of ground elements, and the assumptions in the browsing model, different metrics and variants unfold. In addition to novelty and diversity metrics, we propose different re-ranking strategies to optimize these metrics. Our re-ranking strategies are based on the same item novelty models of their target metrics and apply a re-scoring approach – either direct or greedy – of the outputs of baseline recommendation algorithms by means of a linear combination between the original score of the recommendation and the value of the item novelty model.

The rest of this chapter is organized as follows. In Section 4.2 we describe discovery, relevance and choice in recommendations as the three concepts that build most of our framework. Section 4.3 presents the item novelty models, which are the main component of our unified framework for particularizing the novelty and diversity of the recommendations as the contribution of its composing items. In Section 4.4 we discuss possible browsing models for recommendation lists that provide the schemes that, together with item novelty models, build our metrics. Estimations of the ground models that appeared in the previous sections are proposed in Section 4.5, and the resulting metrics and their equivalences to state-of-the-art proposals are described in Section 4.6. Alternative approaches for modeling novelty and diversity lying outside the generality of our framework appear in Section 4.7. Section 4.8 discuss how to use item novelty models to re-rank the output of baseline recommendation algorithms to optimize novelty and diversity. We report experimental observations validating and illustrating the properties of the proposed metrics and re-ranking strategies in Section 4.9. Finally, Section 4.10 offers the conclusions.

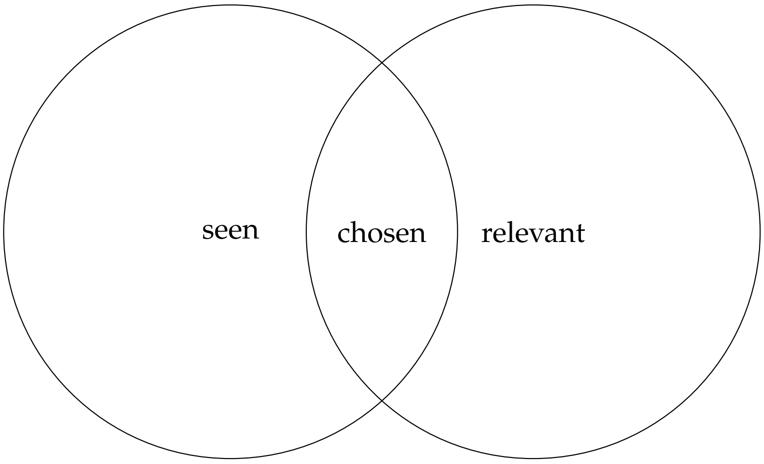

Figure 4.1: Discovery, relevance and choice.

The contents of this chapter have been presented in following published work:

- Vargas, S. and Castells, P. (2011). Rank and relevance in novelty and diversity metrics for recommender systems. In *Proceedings of the 5th ACM Conference on Recommender Systems*, RecSys '11, pages 109–116, New York, NY, USA. ACM
- Castells, P., Vargas, S., and Wang, J. (2011). Novelty and diversity metrics for recommender systems: Choice, discovery and relevance. In *International Workshop on Diversity in Document Retrieval at the 33rd European Conference on Information Retrieval*, DDR'11

4.2 discovery, relevance and choice

The formalization of our framework lies on three fundamental relations between users and items in recommendations:

- Discovery: an item is seen or is known by, or is familiar to a user.
- Relevance: an item is liked, found useful, enjoyed, etc., by a user.
- Choice: an item is used, picked, selected, consumed, bought, etc., by a user.

We assume that these three relations are naturally related: a chosen item has clearly been discovered, and relevant items are more likely to be chosen than irrelevant ones. We further assume, as a simplification, no relation between discovery and relevance – items are discovered independently from their relevance. As a further simplifying assumption, we identify choice with the conjunction of discovery and relevance: an item is chosen if and only if it has been discovered and is relevant, see Figure 4.1.

Conveniently, we may model these three relations as binary random variables over the set of users and the set of items: $\mathrm{seen}, \mathrm{rel}, \mathrm{choose} : \mathcal{U} \times \mathcal{I} \to \{0, 1\}$. Under this probabilistic model, the aforementioned simplifying assumptions can be stated as:

$$p(choose) = p(seen \cap rel) \sim p(seen)\, p(rel)$$

where $choose$ is a shorthand for $choose = 1$, and the same for the other two variables. We are particularly interested in two different probability distributions in which the $seen$, rel and $choose$ random variables are involved:

- **Forced scenario**: expressed as $p(seen\,|\,i)$ (or analogously for rel and $choose$), it measures the probability of a given item i being seen (or liked, or chosen).
- **Free scenario**: expressed as $p(i\,|\,seen)$ (or analogously for rel and $choose$), in which we define the relative probability of an item being the one observed (or liked, or chosen).

Interestingly, these two distributions are equivalent (up to a constant) when we consider uniform item sampling priors $p(i)$:

$$p(i\,|\,seen) = \frac{p(seen\,|\,i)}{\sum_{j \in \mathcal{I}} p(seen\,|\,j)}$$

Despite this equivalence, the use of both distributions is justified by either practical convenience or formal consistency where appropriate.

Discovery, relevance and choice play different roles in our framework. Discovery is used as the basis to define a family of item novelty models. Choice – as the conjunction of discovery and relevance – is used to build models of user browsing behavior over recommendations. Together, browsing models and item novelty models give rise to a fairly wide range of novelty and diversity metrics and variants, as we shall see.

4.3 ITEM NOVELTY MODELS

When assessing novelty and diversity in Recommender Systems, we have to regard – independently from the specific perspective – the general definition of each of them. On the one hand, novelty is understood as the quality of being new or different from what is already known. On the other hand, diversity is a quality of having or being composed of differing elements. These definitions establish a basic distinction between both notions in the sense that the latter (diversity) is exclusively applicable to a set or collection – such as the items in recommendation, the different recommendations issued to the community of users, etc. – while the former (novelty) may be applied more generally but always with respect to some prior experience – the popularity of the recommended items, the items the user already knows, etc. – or, more generally, a certain context. Nevertheless, such distinction does not impose an uncrossable separation between both dimensions for the evaluation of Recommender Systems. In fact, the diversity of a set can be naturally conceived as the aggregate novelty of each of its elements with respect to the others, that is, considering that our novelty-defining context is the rest of the elements of the considered set (or simply, equivalently, the set itself).

Once we see that both novelty and diversity are conceptually related – in particular, that diversity can be treated as a special case of novelty –, we explore in this section the idea that the novelty or the diversity of a recommendation or a set of recommendations can be ultimately decomposed as the aggregation of the individual novelty contributions of each of the items composing the recommendations. Thus, we treat item novelty as the core element in the definition of recommendation novelty and diversity in our framework. We elaborate such item-centric approach to novelty and diversity by proposing the concept of **item novelty models**. As stated in the introduction, our framework is able to describe all the different perspectives of novelty and diversity in recommendations by means these item novelty models. Formally, we define our novelty model as a function $nov : \mathcal{I} \to [0, \infty]$ that assigns higher values the more novel the item is. As we shall see, a fair amount of the metrics proposed in the state of the art can be described by means of a specific item novelty model, and new metrics arise when we consider alternative models.

To unfold an item novelty model, we require two building blocks: a novelty context and a measurement approach. By novelty context, denoted as θ, we denote those aspects or settings involved in the recommendation task that affect the evaluated novelty or diversity perspective of interest. In particular, each different perspective on novelty and diversity corresponds to a distinct context. By measurement approach we mean a particular mathematical basis that helps us quantify the novelty that a item entails. In particular, such measurement approaches are defined upon generic novelty contexts, and will be denoted by $nov(i \,|\, \theta)$. Depending on the information available and the specific novelty or diversity notion, we may be interested in one or other measurement approach. In Section 4.3.2, we present two families for such measurements, one based on the concept of discovery and the other based on the distance between items. A novelty model will be thus formed by the combination of a novelty context – the one of the notion to evaluate – and a measurement approach.

4.3.1 *Novelty Contexts*

As previously introduced, a key component when formalizing item novelty models consist in specifying the particular **context** θ defined by the explored notion of novelty and diversity. Determining accurately the context is key to properly formalizing each particular notion of novelty or diversity. In this section we describe, for the alternative definitions of novelty and diversity in Recommender Systems, the precise context involved in each of them. Table 4.1 summarizes the different novelty contexts for each of the contemplated perspectives on novelty and diversity.

In the **Long Tail Novelty** scenario (see Section 2.3.3.1), we were interested in avoiding recommending a highly reduced set of the most popular items, the so-called short head, and promoting instead recommendations in the more numerous, less popular long tail. The popularity of an item is defined by how many users know about it, and an approximation to such information is readily available in most recommendation scenarios as the user-item interaction data in the form of a rating or play count matrix $\mathcal{R}$. Therefore, we can consider that the context in Long Tail Novelty is the recorded interactions between users and items, that is, our rating matrix $\mathcal{R}$. This represents the available, partial observations of the system of the whole set of interactions between users and items (including those occurring outside the system). Albeit incomplete, these observations are useful to build meaningful estimates of the actual popularity of items.

A related though different notion of novelty considers the **Unexpectedness** (see Section 2.3.3.2) or degree of user-relative unfamiliarity with the provided recommendations. This perspective considers the particular experience of the user to determine the novelty of the recommended items regardless of their popularity. It is clear that the context here is defined by the knowledge of the user u, typically in the form of her profile $\mathcal{I}_u$, that is, the items that the user has interacted with (by listening, watching, buying or consuming them).

A third perspective is the **Temporal Novelty** (see Section 2.3.3.3), in which we consider the effects of issuing recommendations over time to a user. The interest here is in being capable of providing different recommendations over time. In this case, an item is considered novel when it has not been discovered by the user in previous recommendations from the system at hand. Given a recommendation R_u^t for a user u at a point t in time, the context in which we define temporal novelty consists of the previous recommendations that the user received in the past, that is, $\theta = \{R_u^\tau\}_{\tau<t}$.

Another user-oriented perspective consists in the so-called **Intra-List Diversity** (see Section 2.3.3.4), which considers how different are the items in a recommendation between each other. As previously stated, such diversity can be decomposed as the novelty of each item with respect to the others. Clearly, we are in the case that, for each recommended item i in a recommendation R, the context upon which the item novelty is defined as the rest of recommended items $R \setminus \{i\}$.

In the business-oriented perspective, **Sales Diversity** (see Section 2.3.3.5) considers “making the most of the catalog”, i. e., maximizing the exposure of the items of the catalog to avoid concentrating the potential choices of the users around a

Perspective	Context
Long Tail Novelty	$\theta = \mathcal{R}$
Unexpectedness	$\theta = \mathcal{I}_u$
Temporal Novelty	$\theta = \{R_u^\tau\}_{\tau=0}^{t-1}$
Intra-List Diversity	$\theta = R_u$
Sales Diversity	$\theta = \left\{R_v^S\right\}_{v \in \mathcal{U}}$
Sales Novelty	$\theta = \left\{R_u^\Sigma\right\}_{\Sigma \in \mathcal{S}}$

Table 4.1: Summary of the novelty contexts for all the identified novelty and diversity perspectives.

reduced set of the items in the catalog. In such a diversity perspective, we can describe the context of novelty as the set $\left\{R_v^S\right\}_{v \in \mathcal{U}}$ of recommendations provided to the community of users as a whole by a recommender system S.

Finally, the other business-oriented perspective studied is **Sales Novelty** (see Section 2.3.3.6). Under this approach, given a user and a recommender system, we seek to offer recommendations that are novel with respect to the recommendations provided by other recommender systems. In this case, the context is formed by the different recommendations provided by the different systems for the same user u, i. e., $\left\{R_u^\Sigma\right\}_{\Sigma \in \mathcal{S}}$.

4.3.2 *Measurement Approaches*

The other component of our definition of item novelty models consists in what we call **measurement approaches**. These are ways of measuring the novelty provided by an item in a generic context, that is, meta-functions $nov : \mathcal{I} \times \Theta \to [0, \infty]$ that provide numerical values that translate our notion of novelty: the higher the novelty of an item i in a context θ, the higher the value of $nov(i \mid \theta)$ should be. In this section we present two families of measurements which, together to the selection of a novelty context, result in particular instances of novelty models.

4.3.2.1 *Discovery-Based Measurement*

A first set of measurements, the **discovery-based** family, considers the probability of discovery introduced in Section 4.2, in which we calculate how likely is a random variable $seen$ – defined over the pairs of users and items in a sampling space defined by our context – in both free ($p(i \mid seen, \theta)$) and forced ($p(seen \mid i, \theta)$) scenarios. The relation between this probabilistic view of discovery and the novelty is obvious: the higher the probability of an item being known or seen in the sampling space defined by the context θ, the smaller the novelty it adds up. Therefore, we need to provide a mapping between the probability of discovery of an item in the

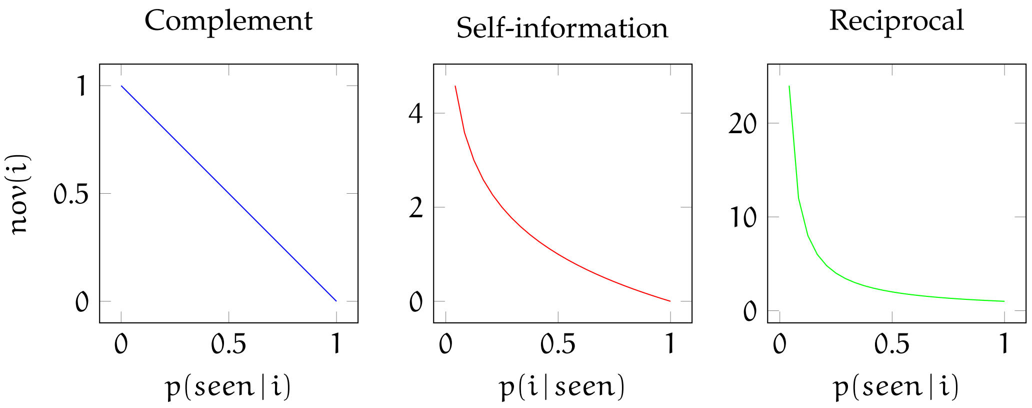

Figure 4.2: The three proposed discovery-based novelty measurement approaches.

form of a decreasing function. We suggest three mappings between the discovery probability and novelty:

- **Complement**: the novelty is the complement of the forced discovery.

  $$\text{nov}^C(i\,|\,\theta) = 1 - p(\text{seen}\,|\,i,\theta)$$

  This is the most straightforward formulation, in which $\text{nov}^C(i\,|\,\theta)$ is simply the probability that the item i has not been seen yet by the user.

- **Self-information**: the novelty is the self information of the free discovery.

  $$\text{nov}^S(i\,|\,\theta) = -\log_2 p(i\,|\,\text{seen},\theta)$$

  This formulation does not have a properly Bayesian interpretation, but it links to a very common way to assess the amount of new information conveyed in a message, as e.g. in Information Theory (Shannon, 1948), and some authors in the area of Recommender Systems (Zhou et al., 2010).

- **Reciprocal**: the novelty is the reciprocal of the forced discovery.

  $$\text{nov}^R(i\,|\,\theta) = \frac{1}{p(\text{seen}\,|\,i,\theta)}$$

  The reciprocal value is a natural and simplest alternative to the complement for a monotonically decreasing function. Even though, again, it does not follow from a pure probabilistic interpretation, it is useful in enabling further connections with widely used diversity metrics in very different fields, as we shall see.

Figure 4.2 provides a comparison of the behavior of these three mappings.

4.3.2.2 *Distance-Based Measurement*

A second family of measurements for item novelty models is motivated by the limitations of the previous one. In particular, the discovery-based approach considers how different an item is from the past experience in terms of a strict boolean identity: an item is novel if it is absent from past experience ($seen = 0$) and not novel otherwise ($seen = 1$). There are reasons however to consider relaxed versions of the boolean view: the knowledge available to the system about what the users have seen is partial – an item might be familiar to a user even if no such interaction has been observed in the system – and, even when a user sees an item for the first time, the resulting novelty gain may range in practice over a gradual rather than a binary scale – that could be the case of sequels or same-themed movies. Thus, we propose an alternative family of measurements, the **distance-based** family, to overcome the limitations of the discovery-based approaches.

In the distance-based approaches, we consider that the novelty of an item is derived from a distance function $dist : \mathcal{I} \times \mathcal{I} \to [0, \infty]$ applied to the measured item and the items appearing in the novelty context. Note that this approach is restricted to novelty contexts consisting on sets of items ($\theta \subset \mathcal{I}$), such as the cases of Unexpectedness and Intra-List Diversity. Given the distances between the assessed item and the items in the context, we can obtain the novelty of the item as an aggregation of the distances, in the form of a **weighted average**:

$$nov^{AD}(i \,|\, \theta) = \sum_{j \in \theta} p(j \,|\, choose, u, \theta) \; dist(i, j)$$

where $p(j \,|\, choose, \theta)$ defines the relative preference of the user for the items in the context. Alternatively, we could also consider the **minimum** of the distances:

$$nov^{MD}(i \,|\, \theta) = \min_{j \in \theta} dist(i, j) \tag{4.1}$$

4.3.3 *Resulting Item Novelty Models*

The combination of contexts and measurement approaches results in specific item novelty models, which form the core of our novelty and diversity framework. We present next a selection of item novelty models of interest (i.e. combinations of context and measurement), either because they appear in the state of the art or because they give rise to novel, interesting metrics further detailed in Section 4.6:

- **Popularity Complement**:

$$nov_{PC}(i) = nov^{C}(i \,|\, \mathcal{R}) = 1 - p(seen \,|\, i, \mathcal{R}) \tag{4.2}$$

- **Free Discovery**:

$$\mathit{nov}_{\mathrm{FD}}(i) = \mathit{nov}^{S}(i \,|\, \mathcal{R}) = -\log_2 p(i \,|\, \mathrm{seen}, \mathcal{R}) \tag{4.3}$$

- **Profile Distance**:

$$\mathit{nov}_{\mathrm{PD}}(i) = \mathit{nov}^{AD}(i \,|\, \mathcal{I}_u) = \sum_{j \in \mathcal{I}_u} p(j \,|\, \mathrm{choose}, u, \mathcal{I}_u)\ \mathrm{dist}(i, j) \tag{4.4}$$

- **Temporal Discovery**:

$$\mathit{nov}_{\mathrm{TD}}(i) = \mathit{nov}^{C}(i \,|\, \{R_u^{\tau}\}_{\tau<t}) = 1 - p(\mathrm{seen} \,|\, i, \{R_u^{\tau}\}_{\tau<t}) \tag{4.5}$$

- **Intra-List Distance**:

$$\begin{aligned}\mathit{nov}_{\mathrm{ILD}}(i) &= \mathit{nov}^{AD}(i \,|\, R_u \setminus \{i\}) \\ &= \sum_{j \in R_u \setminus \{i\}} p(j \,|\, \mathrm{choose}, u, R_u \setminus \{i\})\ \mathrm{dist}(i, j)\end{aligned} \tag{4.6}$$

- **Inter-User Discovery Complement**:

$$\mathit{nov}_{\mathrm{IUDC}}(i) = \mathit{nov}^{C}(i \,|\, \{R_v^S\}_{v \in \mathcal{U}}) = 1 - p(\mathrm{seen} \,|\, i, \{R_v^S\}_{v \in \mathcal{U}}) \tag{4.7}$$

- **Inter-User Reciprocal Discovery**:

$$\mathit{nov}_{\mathrm{IURD}}(i) = \mathit{nov}^{R}(i \,|\, \{R_v^S\}_{v \in \mathcal{U}}) = 1/p(\mathrm{seen} \,|\, i, \{R_v^S\}_{v \in \mathcal{U}}) \tag{4.8}$$

- **Inter-User Free Discovery**:

$$\mathit{nov}_{\mathrm{IUFD}}(i) = \mathit{nov}^{S}(i \,|\, \{R_v^S\}_{v \in \mathcal{U}}) = -\log_2 p(i \,|\, \mathrm{seen}, \{R_v^S\}_{v \in \mathcal{U}}) \tag{4.9}$$

- **Inter-System Discovery Complement**:

$$\mathit{nov}_{\mathrm{ISDC}}(i) = \mathit{nov}^{C}(i \,|\, \{R_u^{\Sigma}\}_{\Sigma \in \mathcal{S}}) = 1 - p(\mathrm{seen} \,|\, i, \{R_u^{\Sigma}\}_{\Sigma \in \mathcal{S}}) \tag{4.10}$$

4.4 BROWSING MODELS

As elaborated on Section 4.3, the aggregation of the individual contributions of the items – in the form of item novelty models – can describe the considered perspectives on novelty and diversity for Recommender Systems. In this section, we study how such aggregation should be performed. In particular, we consider that a recommended item contributes to the evaluated novelty or diversity view inasmuch as it is actually chosen or used by the user, that is, we model the effective or expected contribution of an item i in a recommendation list R_u as $p(choose\,|\,i, R_u)\; nov(i)$. As detailed in Section 4.2, we consider that an item is chosen when it is seen and found relevant. Thus, considering choice in the context of a recommendation results in a user browsing model in which the position and the relevance of the items – among other properties – determine their usage probability. This browsing model is proposed as a solution to one of the detected pitfalls of the current proposals for assessment: lack of position and relevance awareness in the assessment of novelty and diversity.

As thoroughly studied in the area of Information Retrieval (Moffat and Zobel, 2008; Chapelle et al., 2009; Carterette, 2011), the arrangement of each document in a search result list clearly determines the odds of that document being accessed. Equivalently in Recommender Systems, recommended items tend to be arranged in specific spatial patterns – such as lists, grids, pages, etc. – which affect the chances of a recommended item being discovered. The study of such positional bias has been vastly studied in Information Retrieval and Recommender Systems (Herlocker et al., 2004; Hofmann et al., 2014), concretely in the case of search result lists – as displayed in the major Web Search Engines. For illustrative and simplifying purposes, we also consider here the case as rank-awareness in recommendations by assuming list-based displays of our recommendations.

Given a recommendation list R_u for a user u, we consider a browsing model determined by the probability distribution that a user chooses each item in the recommendation. As assumed in Section 4.2, this choosing probability is determined by the discovery probability – related to the ranking position in the item – and the user-perceived relevance:

$$p(choose\,|\,i, R_u) = p(seen\,|\,i, u, R_u)\; p(rel\,|\,i, u)$$

The component $p(seen\,|\,i, u, R_u)$ represents the probability that the target user will actually see the item i when he is browsing the ranked list R_u. This component allows for the introduction of a rank discount by having $p(seen\,|\,i, u, R_u)$ reflect the fact that the lower an item is ranked in R_u, the less likely it will be seen. A realistic model may take into consideration that users eventually get tired of browsing, or get satisfied by enough items, or a combination of both, and stop browsing at some point before the end of the list, leaving a number of recommended items unread – which would play no part in the effective recommendation novelty the user will perceive.

In general we assume a so-called cascade model (Clarke et al., 2008) where the user browses the items by ranking order without jumps, until she stops. At each position k in the ranking, the user makes a decision whether or not to continue, which we model as a binary random variable $cont$, where $p(cont\,|\,k, u, R_u)$ is the probability that the user decides to continue browsing the next item at position $k+1$. With this scheme we have, by recursion:

$$\begin{aligned} p(\text{seen}\,|\,i_k, u, R_u) &= p(\text{seen}\,|\,i_{k-1}, u, R_u)\; p(\text{cont}\,|\,k-1, u, R_u) \\ &= \prod_{l=1}^{k-1} p(\text{cont}\,|\,l, u, R_u) \end{aligned}$$

where i_k is the document ranked at position k.

Now there are several ways – of varying complexity – in which $p(\text{cont}\,|\,l, u, R_u)$ can be modeled. A simple one is to consider a constant $p(\text{cont}\,|\,l, u, R_u) = \beta$, whereby we get an exponential discount $p(\text{seen}\,|\,i_k, u, R_u) = \beta^{k-1}$. This is the approach taken in the Rank-Biased Precision (RBP) search performance metric (Moffat and Zobel, 2008). We may consider instead that the user will stop as soon as – and only when – she finds the first item of her taste. In that case, the discount is $p(\text{seen}\,|\,i_k, u, R_u) = \prod_{l=1}^{k-1} (1 - p(\text{rel}\,|\,i_l, u))$, similar to the ERR metric (Chapelle et al., 2009), or the models in (Radlinski et al., 2008). We might consider more complex and general models, such as:

$$p(\text{seen}\,|\,i_k, u, R_u) = p(\text{cont}|\neg\text{rel})^{k-1} \prod_{l=1}^{k-1} (1 - p(\text{rel}\,|\,i_l, u))$$

similar to (Clarke et al., 2008), or

$$p(\text{cont}\,|\,l, u, R) = p(\text{cont}\,|\,\text{rel})\; p(\text{rel}\,|\,i_l, u) + p(\text{cont}\,|\,\neg\text{rel})\;\; (1 - p(\text{rel}\,|\,i_l, u))$$

and so forth. In general, we suggest using a decreasing rank discount function:

$$p(\text{seen}\,|\,i, u, R_u) = \text{disc}(k_i)$$

where k_i is the rank of item i. Such ranking function can be chosen as it suits, either based on the aforementioned cascade model or heuristic ones, such as a logarithmic discount as in the normalized Discount Cumulative Gain (nDCG) (Järvelin and Kekäläinen, 2000), a Zipfian discount, etc., or even no discount by $\text{disc}(k) = 1$, as if the user always browsed the whole list.

Finally, the combination of the novelty models from Section 4.3 and the browsing model presented in the present section results in the following metric schemes for user-oriented metrics:

$$m(R_u) = C_u \sum_{i \in R_u} \text{disc}(k_i)\; p(\text{rel}\,|\,i, u)\; \text{nov}(i)$$

In the case of business-oriented metrics, which do not provide scores for individual recommendations, we formulate our metric scheme as follows:

$$m(S) = C_S \sum_{u \in \mathcal{U}} C_u \sum_{i \in R_u^S} \mathrm{disc}(k_i)\; p(\mathrm{rel} \,|\, i, u)\; \mathrm{nov}(i)$$

where C_S and C_u are system and user normalization terms. Such normalization terms usually serve as a means to properly compare or average different recommendation measurements. In the case of the user normalization term C_u, typical choices are:

- Expected browsing depth, as in (Moffat and Zobel, 2008) and discussed in (Clarke et al., 2011a), computed as

$$\begin{aligned} \frac{1}{C_u} &= \sum_{i \in R_u} k_i\; p(\mathrm{seen} \,|\, i, u, R_u)\;\; (1 - p(\mathrm{cont} \,|\, k_i, u, R_u)) \\ &= \sum_{i \in R_u} k_i\;\; (\mathrm{disc}(k_i) - \mathrm{disc}(k_i - 1)) \\ &= \sum_{i \in R_u} \mathrm{disc}(k_i) \end{aligned}$$

- Ideal novelty, as in nDCG and α-nDCG (Clarke et al., 2008), which is the novelty achieved an ideal ranking.

Note that in case of no ranking discount, i.e. $\mathrm{disc}(k) = 1$, our user-metric scheme with the expected browsing depth would result in a simpler relevance-weighted average novelty ($C_u = 1/|R_u|$). As for the case of business-oriented metrics, a typical normalization term C_S would simply be the number of evaluated users $C_S = 1/|\mathcal{U}|$.

4.5 Estimation of Ground Models

In the previous sections formal models for item novelty and recommendation list browsing models were developed, resulting in two general metric schemes for user and business-metrics for novelty and diversity. In order to make this scheme fully implementable, we need to provide practical methods to estimate the models – discovery, distance and relevance – upon which we have built the framework.

4.5.1 *Discovery*

Discovery plays a important role throughout the framework, either as the basis for a definition of novelty measurements or as a rank-aware discount for our browsing models. The latter case was already extensively discussed in Section 4.4, where we suggested several elaborated models for cascade-based browsing and a simple ranking discount component $\mathrm{disc}(k)$. In this section, we provide practical estimations for the discovery models used in Section 4.3.3.

Starting with the novelty models for **Long Tail Novelty** (Equations 4.2 and 4.3), we can take the sampling space of the observed pairs in the user-item interactions matrix $\mathcal{R}$ to estimate the popularity of a given item in both the forced and the free discovery distributions:

$$p(\text{seen}\,|\,i,\mathcal{R}) \sim \frac{|\mathcal{U}_i|}{|\mathcal{U}|} \tag{4.11}$$

$$p(i\,|\,\text{seen},\mathcal{R}) \sim \frac{|\mathcal{U}_i|}{\sum_{j\in\mathcal{I}}|\mathcal{U}_j|} = \frac{|\mathcal{U}_i|}{|\mathcal{R}|} \tag{4.12}$$

As indicated in Section 4.2, these two distributions are proportional and for most purposes equivalent. More elaborated models can take advantage of the specific type of interactions between users and items – ratings, play counts, logged accesses, etc – but, from our experience, the previous simple estimations are adequate for modeling the popularity of the items.

We follow by dealing with the case of **Temporal Discovery** (Equation 4.5). In this case, the context is composed of the previous recommendations that the user received. A simple initial approach would consist in a binary criterion in which an item that has not been seen in previous recommendations is considered completely novel and not novel otherwise:

$$p(\text{seen}\,|\,i,\{R_u^\tau\}_{\tau<t}) \sim \mathbb{1}_{\exists\tau<t\,:\,i\in R_u^\tau} \tag{4.13}$$

However, such a simple model does not take into account the browsing model involved in the previous recommendations. In fact, we could consider the case of an item that appeared in a previous recommendation but was not seen by the user: in that case assuming no novelty would not be adequate to the real experience. We can alleviate such defect by incorporating a ranking discount model such as those defined in Section 4.4. This way, if we compare our present recommendation with the previous one, we can model a rank-aware temporal novelty as:

$$p(\text{seen}\,|\,i,R_u^{t-1}) \sim \text{disc}(k_i^{R_u^{t-1}}) \tag{4.14}$$

where $k_i^{R_u^{t-1}}$ is the position of item i in the list R_u^{t-1}. This last estimation can be generalized to consider more previous recommendations and the effect time in the discovery process – a user may forget the recommendations received a while ago.

When we consider the **Inter-User Discovery** (Equations 4.7, 4.8 and 4.9), that is, the discovery with respect to the recommendations issued to a community of users by a particular recommender system, we can initially take, as in the Long Tail Novelty case, a simple frequentist approach:

$$p(\text{seen}\,|\,i,\{R_v\}_{v\in\mathcal{U}}) \sim \frac{|\{v\in\mathcal{U}\,:\,i\in R_v\}|}{|\mathcal{U}|} \tag{4.15}$$

$$p(i \,|\, \text{seen}, \{R_v\}_{v \in \mathcal{U}}) \sim \frac{|\{v \in \mathcal{U} : i \in R_v\}|}{\sum_{v \in \mathcal{U}} |R_v|} \tag{4.16}$$

However, as in the case of temporal diversity, it seems reasonable to consider the actual discovery of the user in her recommendation:

$$p(\text{seen} \,|\, i, \{R_v\}_{v \in \mathcal{U}}) \sim \frac{\sum_{v \in \mathcal{U}} \text{disc}(k_i^{R_v})}{|\mathcal{U}|} \tag{4.17}$$

$$p(i \,|\, \text{seen}, \{R_v\}_{v \in \mathcal{U}}) \sim \frac{\sum_{v \in \mathcal{U}} \text{disc}(k_i^{R_v})}{\sum_{j \in \mathcal{I}} \sum_{v \in \mathcal{U}} \text{disc}(k_j^{R_v})} \tag{4.18}$$

Finally, the estimation of the **Inter-System Discovery** (Equation 4.10) can take a similar form to that of Inter-User Discovery:

$$p(\text{seen} \,|\, i, \left\{R_u^{\Sigma}\right\}_{\Sigma \in \mathcal{S}}) \sim \frac{\left|\left\{\Sigma \in \mathcal{S} : i \in R_u^{\Sigma}\right\}\right|}{|\mathcal{S}|} \tag{4.19}$$

and also consider the rank of the items in the recommendations:

$$p(\text{seen} \,|\, i, \left\{R_u^{\Sigma}\right\}_{\Sigma \in \mathcal{S}}) \sim \frac{\sum_{\Sigma \in \mathcal{S}} \text{disc}(k_i^{R_u^{\Sigma}})}{|\mathcal{S}|} \tag{4.20}$$

4.5.2 *Distance*

The distance-based novelty models defined in Equations 4.4 and 4.6 – Profile Distance and Intra-List Diversity – were left without giving further details about how to estimate the distance between items and the specific of the relative probability $p(j \,|\, \text{choose}, u, \theta)$ of choice of the items $j \in \theta$ in the context. In this section we specify different possibilities for their instantiation.

The Recommender Systems community has dedicated a good deal of effort in proposing similarity measures between items as an integral part of many recommendation algorithms. We build upon this concept of similarity by defining distance as the complement of similarity:

$$\text{dist}(i, j) = 1 - \text{sim}(i, j)$$

where $\text{sim} : \mathcal{I} \times \mathcal{I} \to [0, 1]$ is a normalized similarity function. Different choices for the definition of similarities between items have been proposed, being distinguished the following two main families:

- Content-based: two items are similar if they share similar intrinsic features; this is the approach taken in (Ziegler et al., 2005).
- Collaborative filtering-based: two items are similar if many users interacted with both of them, as seen in (Ribeiro et al., 2012; Zhang et al., 2012).

These two approaches are not necessarily equivalent. In fact, two items with similar characteristics may not have common users and two items with common users do not have to be similar in content – although intuition may give a higher probability for the latter.

In the content-based case, given a set of features $\mathcal{F}$ that captures the content-based properties of our items, we can compute their similarity by means of well-known similarity coefficients, such as the cosine:

$$\mathrm{sim}_{\mathcal{F},\mathrm{cosine}}(i,j) = \frac{|\mathcal{F}_i \cap \mathcal{F}_j|}{\sqrt{|\mathcal{F}_i|\,|\mathcal{F}_j|}} \tag{4.21}$$

Similarly, we can use an equivalent definition for the collaborative filtering-based case by changing the features by the users that interacted with the items (Cremonesi et al., 2010; Aiolli, 2013):

$$\mathrm{sim}_{\mathrm{CF},\mathrm{cosine}}(i,j) = \frac{|\mathcal{U}_i \cap \mathcal{U}_j|}{\sqrt{|\mathcal{U}_i|\,|\mathcal{U}_j|}} \tag{4.22}$$

Alternatively, we can use different similarity coefficients such as Jaccard or Dice and take advantage of the specific form of the interactions between users and items such as rating data, play counts, etc.

For the estimation of the relative choice of the **Profile Distance** (Equation 4.4), we opt for a relevance-only based approach:

$$p(j\,|\,\mathrm{choose}, u, \mathcal{I}_u) \sim \frac{p(\mathrm{rel}\,|\,j,u)}{\sum_{j' \in \mathcal{I}_u} p(\mathrm{rel}\,|\,j',u)} \tag{4.23}$$

In this case, we could also complement this estimation by adding a discovery-based approach to take into account how recently the user interacted with the item j – older items may have been forgotten.

In the case of the **Intra-List Distance** novelty model (Equation 4.6), whose context is composed by the recommendation itself, taking into account an adaptation of the browsing model in Section 4.4 results in the following relative choice probability:

$$p(j\,|\,\mathrm{choose}, u, R_u \setminus \{i\}) \sim \frac{\mathrm{disc}(k_j\,|\,k_i)\; p(\mathrm{rel}\,|\,j,u)}{\sum_{j' \in R_u \setminus \{i\}} \mathrm{disc}(k_{j'}\,|\,k_i)\; p(\mathrm{rel}\,|\,j',u)} \tag{4.24}$$

where $\mathrm{disc}(k_j\,|\,k_i) = \mathrm{disc}(\max(0, k_j - k_i))$ is a relative ranking discount to consider that the items ranked before the target item have – by definition of the browsing model – always been discovered.

4.5.3 *Relevance*

Lastly, we seek ways of estimating the probability of relevance $p(\mathrm{rel}\,|\,i,u)$ that models the preferences of the user. As is standard in the evaluation of recommender systems, we assume a partition of the interaction data $\mathcal{R}$ into a training and a test set: $\mathcal{R} = \mathcal{R}_{\mathrm{train}} \uplus \mathcal{R}_{\mathrm{test}}$. Usually, recommendations are generated by using only the data from training, and test data play the role of relevance judgments in the Cranfield Information Retrieval evaluation methodology. This way, for a given item in a recommendation, we can have three different situations with respect to the test data:

- The item appears in the test set of the user and is judged as relevant – high rating, like, etc.

- The item appears in the test set of the user and is judged as irrelevant – low rating, dislike, etc.

- The item does not appear in the test set of the user.

In relevance-oriented evaluation methodologies, the second and the third cases are commonly treated equally as non relevant items. While this assumption has a reasonable basis – most items are irrelevant to users and the test set is a good lower bound of the interests of the user –, it may become impractical in our novelty or diversity-oriented scenario. In particular, we have observed that, in a strict relevance-aware evaluation, most of the recommended items are judged as irrelevant and therefore they do not contribute to the novelty or diversity of the recommendation. This can cause a strong bias of our metrics towards accurate but marginally novel or diverse recommendations.

Our proposal for overcoming the sparsity in the relevance judgments extracted from the test set consists in a simple background or default relevance probability assigned to items absent in the test:

$$p(\mathrm{rel}\,|\,i,u) \sim \begin{cases} 1 & (u,i) \in \mathcal{R}_{\mathrm{test}} \wedge r_{ui} \geqslant \rho \\ 0 & (u,i) \in \mathcal{R}_{\mathrm{test}} \wedge r_{ui} < \rho \\ b & (u,i) \notin \mathcal{R}_{\mathrm{test}} \end{cases} \tag{4.25}$$

where β is the background probability of an unobserved interaction being relevant and ρ is a threshold for the values in $\mathcal{R}$ that determines whether the relevance judgment is positive or negative. In this case, we set a global probability b that a item without any recorded interaction is found relevant by the user. Such probability could be estimated, e.g., by conducting user studies to estimate the probability of random items to be found relevant in a recommendation scenario. Further, we could consider user or item-dependent background probabilities to take into account the particularities of the users – demanding users would have a lower probability b than permissive users – or the items – items liked by different types of users may have higher chances of being relevant than niche items. In our experiments in Section 4.9, we opt for setting a heuristic value for b that shows a good empirical trade-off between novelty or diversity and relevance.

More elaborated relevance models would take advantage of graded relevance. That would be the case, for example, of taking the probability of relevance as done by Chapelle et al. (2009):

$$p(rel\,|\,i,u) \sim \frac{2^{g(u,i)} - 1}{2^{g_{max}}}$$

where $g(u,i) = \max\left(0, r(u,i) - \rho\right)$ is a threshold function for the graded relevance.

4.6 RESULTING METRICS

Once we have defined the novelty and browsing models and some practical estimations of the ground models they are based on, we are in a position to enunciate several metrics that unfold from our framework.

Starting with the **Long Tail Novelty**, plugging the Popularity Complement Novelty (Equations 4.2 and 4.11) in our metric scheme results in the **Expected Popularity Complement** (EPC) metric:

$$EPC(R_u) = C_u \sum_{i \in R_u} disc(k_i)\; p(rel\,|\,i,u)\; \left(1 - \frac{|\mathcal{U}_i|}{|\mathcal{U}|}\right) \tag{4.26}$$

Similarly, taking the Free Discovery model (Equations 4.3 and 4.12) we get the **Expected Free Discovery** (EFD) metric:

$$EFD(R_u) = -C_u \sum_{i \in R_u} disc(k_i)\; p(rel\,|\,i,u)\; \log_2 \frac{|\mathcal{U}_i|}{|\mathcal{R}|} \tag{4.27}$$

This last metric is equivalent to the Mean Inverse User Frequency (MIUF, see Equation 2.1) defined in (Zhou et al., 2010) by discarding any rank or relevance components:

$$\begin{aligned} EFD(R) &= -\frac{1}{|R|} \sum_{i \in R} \log_2 \frac{|\mathcal{U}_i|}{|\mathcal{R}|} \\ &= MIUF(R) + \log_2 |\mathcal{R}| - \log_2 |\mathcal{U}| \end{aligned}$$

Turning to the notion of **Unexpectedness**, we apply the generic browsing model to the Profile Distance Novelty (Equations 4.4 and 4.23) to get the **Expected Profile Distance** (EPD) metric:

$$\mathrm{EPD}(R_u) = C'_u \sum_{i \in R_u} \sum_{j \in \mathcal{I}_u} \mathrm{disc}(k_i)\ p(\mathrm{rel}\,|\,i,u)\ p(\mathrm{rel}\,|\,j,u)\ \mathrm{dist}(i,j) \tag{4.28}$$

where $C'_u = C_u / \sum_{j' \in \mathcal{I}_u} p(\mathrm{rel}\,|\,j',u)$. Our metric resembles that of Adamopoulos and Tuzhilin (in press) in Equation 2.2 by discarding rank and relevance, assuming that the set of expected or obvious recommendations of user u are the items in her profile ($E_u = \mathcal{I}_u$) and taking the average distance as the model for distance of one recommended item to the expected set ($\mathrm{dist}(i, \mathcal{I}_u) = \frac{1}{|\mathcal{I}_u|} \sum_{j \in \mathcal{I}_u} \mathrm{dist}(i,j)$):

$$\begin{aligned} \mathrm{EPD}(R) &= \frac{1}{|R|\ |\mathcal{I}_u|} \sum_{i \in R} \sum_{j \in \mathcal{I}_u} \mathrm{dist}(i,j) \\ &= \mathrm{Unexp}_3(R) \end{aligned}$$

Our proposed metric for **Intra-List Diversity** results from using a fully rank and relevance-aware modeling for both the Intra-List Distance (Equations 4.6 and 4.24) and the browsing, which we call the **Expected Intra-List Distance** (EILD):

$$\mathrm{EILD}(R_u) = \sum_{i,j \in R_u} C_i\ \mathrm{disc}(k_i)\ \mathrm{disc}(k_j\,|\,k_i)\ p(\mathrm{rel}\,|\,i,u)\ p(\mathrm{rel}\,|\,j,u)\ \mathrm{dist}(i,j) \tag{4.29}$$

where $C_i = C_u / \sum_{j' \in R_u \setminus \{i\}} \mathrm{disc}(k_{j'}\,|\,k_i)\ p(\mathrm{rel}\,|\,j',u)$. Our metric generalizes the Intra-List Distance (ILD, see Equation 2.4) of Smyth and McClave (2001):

$$\begin{aligned} \mathrm{EILD}(R) &= \frac{1}{|R|\,(|R|-1)} \sum_{i,j \in R} \mathrm{dist}(i,j) \\ &= \mathrm{ILD}(R) \end{aligned}$$

Note that the previous distance-based metrics (EPD and EILD) can be used with both similarity-derived distances discussed in Section 4.5.2: content-based and collaborative filtering. Despite the common basis, we expect the specific instances of such metrics to be mostly unrelated, given that the similarity principles behind them are not connected.

In our framework, the Temporal Novelty metric (TN, see Equation 2.3) introduced by Lathia et al. (2010):

$$\mathrm{TN}(R^t_u) = \frac{\left|R^t_u \setminus R^{t-1}_u\right|}{|R^t_u|}$$

can be seen as a particular case of our **Expected Temporal Discovery** (ETD) reduced to consider the previous recommendation in the Temporal Discovery model (Equations 4.5 and 4.14):

$$\mathrm{ETD}(R_u^t) = C_u \sum_{i \in R_u^t} \mathrm{disc}(k_i^{R_u^t})\, p(\mathrm{rel}\,|\,i,u)\, (1 - \mathrm{disc}(k_i^{R_u^{t-1}})) \tag{4.30}$$

Our **Sales Diversity** novelty models connect with several metrics found in the state of the art for evaluating this notion of diversity. In particular, when taking the Inter-User Discovery Complement (Equations 4.7 and 4.17) we get the **Expected Inter-User Discovery Complement** (EIUDC) metric:

$$\mathrm{EIUDC}(S) = C_S \sum_{u \in \mathcal{U}} C_u \sum_{i \in R_u} \mathrm{disc}(k_i^{R_u})\, p(\mathrm{rel}\,|\,i,u) \left(1 - \frac{\sum_{v \in \mathcal{U}} \mathrm{disc}(k_i^{R_v})}{|\mathcal{U}|}\right) \tag{4.31}$$

which turns out to be the Inter-User Diversity (IUD, see Equation 2.9) of Bellogín et al. (2010) and Zhou et al. (2010) if we discard rank and position biases:

$$\begin{aligned}\mathrm{EIUDC}(S) &= \frac{1}{|\mathcal{U}|} \sum_{u \in \mathcal{U}} \frac{1}{|R_u|} \sum_{i \in R_u} 1 - \frac{|\{v \in \mathcal{U} : i \in R_v\}|}{|\mathcal{U}|} \\ &= \frac{1}{|\mathcal{U}|^2} \sum_{u,v \in \mathcal{U}} \frac{|R_u \setminus R_v|}{|R_u|} = \frac{|\mathcal{U}| - 1}{|\mathcal{U}|}\, \mathrm{IUD}(S)\end{aligned}$$

If we further assume a fixed recommendation list size, i. e., $\forall v \in \mathcal{U}\ |R_v| = N$, EIUDC and IUD turn out to be equivalent to the Gini-Simpson Index (GSI), which is the probability of randomly picking two different items from the set of recommendations:

$$\begin{aligned}\mathrm{IUD}(S) &= \frac{1}{|\mathcal{U}|\,(|\mathcal{U}|-1)\,N} \sum_{u,v \in \mathcal{U}} |R_u \setminus R_v| = \frac{1}{|\mathcal{U}|\,(|\mathcal{U}|-1)\,N} \sum_{u,v \in \mathcal{U}} N - |R_u \cap R_v| \\ &= 1 - \frac{1}{|\mathcal{U}|\,(|\mathcal{U}|-1)\,N} \sum_{u,v \in \mathcal{U}} |R_u \cap R_v| \\ &= 1 - \frac{1}{|\mathcal{U}|\,(|\mathcal{U}|-1)\,N} \sum_{i \in \mathcal{I}} |\{u \in \mathcal{U} : i \in R_u\}|^2 \\ &= 1 + \frac{|\mathcal{U}|\ N}{|\mathcal{U}|-1} \sum_{i \in \mathcal{I}} \left(\frac{|\{u \in \mathcal{U} : i \in R_u\}|}{|\mathcal{U}|\ N}\right)^2 \\ &= 1 + \frac{|\mathcal{U}|\ N}{|\mathcal{U}|-1} \sum_{i \in \mathcal{I}} p(i\,|\,S)^2 = \frac{|\mathcal{U}|\ N}{|\mathcal{U}|-1} (\mathrm{GSI}(S) - 1) + 1\end{aligned}$$

where $p(i\,|\,S)$ is the probability of a recommended item being drawn from the recommendation lists generated by a system S, as seen in Equation 2.7. Analogously, if we take the Inter-User Reciprocal Discovery (Equations 4.8 and 4.17) we get the **Expected Inter-User Reciprocal Discovery** (EIURD):

$$\mathrm{EIURD}(S) = C_S \sum_{u \in \mathcal{U}} C_u \sum_{i \in R_u} \mathrm{disc}(k_i^{R_u})\, p(rel \,|\, i, u)\, \frac{|\mathcal{U}|}{\sum_{v \in \mathcal{U}} \mathrm{disc}(k_i^{R_v})} \tag{4.32}$$

whose rank and relevance-unaware version is equivalent to the Aggregate Diversity (see Equation 2.5) of Adomavicius and Kwon (2012) assuming a fixed recommendation list size N:

$$\begin{aligned} \mathrm{EIURD}(S) &= \frac{1}{|\mathcal{U}|} \sum_{u \in \mathcal{U}} \frac{1}{N} \sum_{i \in R_u} \frac{|\mathcal{U}|}{|\{v \in \mathcal{U} : i \in R_v\}|} \\ &= \frac{1}{N} \sum_{i \in \bigcup_v R_v} \frac{1}{|\{v \in \mathcal{U} : i \in R_v\}|} \sum_{u \,:\, i \in R_u} 1 \\ &= \frac{1}{N} \left| \bigcup_{v \in \mathcal{U}} R_v \right| = \frac{1}{N} \,\mathrm{Aggr\text{-}div}(S) \end{aligned}$$

Also, the metric derived from the Inter-User Free Discovery model (Equations 4.9 and 4.18), which we call **Expected Inter-User Free Discovery** (EIUFD):

$$\mathrm{EIUFD}(S) = C_S \sum_{u \in \mathcal{U}} C_u \sum_{i \in R_u} \mathrm{disc}(k_i^{R_u})\, p(rel \,|\, i, u)\, \log_2 \frac{\sum_{v \in \mathcal{U}} \mathrm{disc}(k_i^{R_v})}{\sum_{j \in \mathcal{I}} \sum_{v \in \mathcal{U}} \mathrm{disc}(k_j^{R_v})} \tag{4.33}$$

is equivalent without rank and relevance to using the Entropy (Patil and Taillie, 1982) (see Equation 2.8) with a fixed recommendation list size N:

$$\begin{aligned} \mathrm{EIUFD}(S) &= \frac{1}{|\mathcal{U}|} \sum_{u \in \mathcal{U}} \frac{1}{N} \sum_{i \in R_u} \log_2 \frac{|v \in \mathcal{U} : i \in R_v|}{|\mathcal{U}|\ N} \\ &= \sum_{i \in \mathcal{I}} \frac{|v \in \mathcal{U} : i \in R_v|}{|\mathcal{U}|\ N} \log_2 \frac{|v \in \mathcal{U} : i \in R_v|}{|\mathcal{U}|\ N} \\ &= \sum_{i \in \mathcal{I}} p(i \,|\, S)\ \log_2 p(i \,|\, S) = H(S) \end{aligned}$$

Sales novelty, finally, is covered by our Inter-System Discovery Complement (Equations 4.10 and 4.20) with the Expected Inter-System Discovery Complement (EISDC):

$$\mathrm{EISDC}(R_u) = C_u \sum_{i \in R_u} \mathrm{disc}(k_i)\, p(rel \,|\, i, u) \left(1 - \frac{\sum_{\Sigma \in \mathcal{S}} \mathrm{disc}(k_i^{R_u^{\Sigma}})}{|\mathcal{S}|} \right) \tag{4.34}$$

Unsurprisingly, this metric is a generalization of the Inter-System Diversity (ISD, see Equation 2.10) of Bellogín et al. (2010):

$$\begin{aligned}\mathrm{EISDC}(R_u^S) &= \frac{1}{|R_u|}\sum_{i\in R_u} 1 - \frac{\left|\{\Sigma \in \mathcal{S} : i \in R_u^\Sigma\}\right|}{|\mathcal{S}|} \\ &= \frac{1}{|\mathcal{S}|}\sum_{\Sigma\in\mathcal{S}} \frac{\left|R_u^S \setminus R_u^\Sigma\right|}{|R_u^S|} = \frac{|\mathcal{S}|-1}{|\mathcal{S}|}\ \mathrm{ISD}(R_u^S)\end{aligned}$$

4.7 FURTHER GENERALIZATION

As illustrated in the previous section, our framework generalizes many of the approaches for assessing novelty and diversity in the field. However, we also found the case of other proposals which do not fall under the generality of our framework. Some of them are loosely connected to our proposals as they consider a notion of item or document novelty, while others consider properties of recommendations as a whole that cannot be particularized by the individual contributions of the items composing them. In this section we succinctly present some of these alternative approaches to model novelty and diversity of the recommendations outside our framework.

4.7.1 *Alternative Aggregation of Item Novelty*

The problem of Sales Diversity has been treated outside the area of Recommender Systems given its interest in diverse field such as Ecology, Economics or Marketing. Consequently, there are many alternative choices for assessing the diversity of a population, a market, etc. Entropy and the Gini Simpson Index, as seen in the previous section, can be explained by means of our framework by considering an inter-user free discovery model. There are other approaches that, however, cannot be characterized by means of our framework, since they do not take a browsing approach for aggregating the novelty of each item. That is the case of the **Gini Index**, which was initially conceived to represent the income distribution of a population, and can also be applied to measure sales concentration (Fleder and Hosanagar, 2009). Given an income distribution, the Gini Index is defined as the area between the Lorentz curve of the distribution and the Lorentz curve of a perfectly equal distribution. Considering the "income" of an item as the discovery of an item with respect to the recommendations issued to the users, the Gini index of the items can be computed as:

$$\mathrm{Gini}(S) = \frac{1}{|\mathcal{I}|-1}\sum_{k=1}^{|\mathcal{I}|}(2k-|\mathcal{I}|-1)\ p(i_k\,|\,\mathrm{seen}, \{R_v\}_{v\in\mathcal{U}})$$

where i_k is the item with the k-th lowest income (discovery probability). This index takes values between 0 and 1, giving lower values for higher equality in the "income" distribution of the items – a value of 0 means that all the items appear in the same number of recommendations, while a value of 1 means that all recommendations are composed of the same item.

4.7.2 *Intent-Aware approaches*

The most common and best-known approach for evaluating the diversity of search results is the so called **Intent-Aware** framework (see Section 2.4). Given a query q for which a list of documents R is retrieved, the Intent-Aware framework considers a set of subtopics s – possible interpretations or facets of the query – and calculates the marginal relevance of the result list R for each of them:

$$\text{M-IA}(R_q) = \sum_s p(s \mid q)\; m(R_q \mid s) \tag{4.35}$$

where $m(R \mid s)$ is the relevance of the result list R with respect to subtopic s. A usual choice for the relevance metric m is the Expected Reciprocal Rank (ERR) of Chapelle et al. (2009), which in its intent-aware variant ERR-IA (see Equation 2.12) can be formulated as:

$$\begin{aligned}\text{ERR-IA}(R_q) &= \sum_s p(s \mid q) \sum_{k=1}^{|R_q|} \frac{1}{k}\, p(rel \mid d_k, s) \prod_{j=1}^{k-1} \big(1 - p(rel \mid d_j, s)\big)\\ &= \sum_{d \in R_q} \frac{1}{k_d} \sum_s p(s \mid q)\; p(rel \mid d, s) \prod_{d' \,:\, k_{d'} < k_d} \big(1 - p(rel \mid d', s)\big)\end{aligned}$$

where d_k is the document ranked at position k in R_q and $p(rel \mid d_k, s)$ is the probability of relevance of document d_k with respect to subtopic s. The last reformulation of ERR-IA suggest that this metric can be described in terms of a document (instead of item) novelty model. In particular, if we consider the following metric scheme derived from our browsing model without considering relevance:

$$m(R) = \sum_{d \in R} disc(k_d)\; nov(i)$$

we can see that ERR-IA metric results from choosing a reciprocal discount model $disc(k) = 1/k$ and the following document novelty model:

$$nov_{\text{ERR-IA}}(d \mid R) = \sum_s p(s \mid q)\; p(rel \mid d, s) \prod_{d' \,:\, k_{d'} < k_d} \big(1 - p(rel \mid d', s)\big)$$

Similarly, another frequent choice for evaluating search result diversity, α-nDCG (see Equation 2.13), results from the previous scheme by taking an ideal normalization for C_u, a logarithmic rank discount $\mathrm{disc}(k) = 1/\log_2(k+1)$ and the following novelty model:

$$\mathrm{nov}_{\alpha\text{-nDCG}}(d\,|\,R) = \sum_s \mathrm{rel}(d,s) \prod_{d'\,:\,k_{d'}<k_d} (1-\alpha\,\mathrm{rel}(d',s)) \tag{4.36}$$

where $\mathrm{rel}(d,s)$ is a binary relevance judgment of document d with respect to the subtopic s and α a redundancy parameter which control the redundancy of documents covering previously covered subtopics.

As we can see, these approaches to Intra-List Diversity clearly differ from our distance-based proposed Expected Intra-List Distance. Although both take the rest of retrieved documents or items as the novelty context, the Intent-Aware metrics only consider the previously ranked documents and take a subtopic marginal relevance-based measurement approach.

Adapting this Intent-Aware framework for assessing the diversity of recommendations is an interesting and promising line of work that allows connecting perspectives of Recommender Systems and Information Retrieval. We explore such connections in Chapter 5 by establishing analogies between the search and recommendations problems and making intent-oriented proposals for enhancing the diversity in recommendations.

4.7.3 *Beyond Item Novelty*

However general the item novelty modeling, there are a handful of approaches that are outside of the scope of particularizing the novelty of the recommendations by the individual contributions of the items. All these approaches share a subtopic-oriented approach, that is, there is a certain division or categorization of the interests of the user or query, and the diversity of the metrics is evaluated in terms of the coverage of these interests.

One of the earliest approaches is the one taken by Zhai et al. (2003) (see Section 2.4.1.1) in which they consider the amount of subtopics covered by a search result list by means of the **Subtopic Recall** (S-recall, see Equation 2.11) metric:

$$\text{S-}\mathrm{recall}(R_q) = \frac{\left|\bigcup_{d\in R_q} \mathrm{subtopics}_d\right|}{n_{\mathrm{subtopics}}} \tag{4.37}$$

A recent approach by Dang and Croft (2012), which we call the **Proportionality Framework** (see Section 2.4.1.4), compares the proportion of documents covering each subtopic in result list with the expected one. The authors propose evaluation and enhancement approaches inspired on a seat assignment system for legislative elections in some countries.

Finally, such subtopic-oriented approaches can also be considered in the recommendation domain, specially when the interests of the user can be classified in particular categories, as in the case of genres for movie, music or book domains. In particular, we present in Chapter 6 a proposal for modeling genre-based diversity in recommendation list by considering coverage, redundancy and size-awareness as the three main requirements. We propose the **Binomial framework** to satisfy these three properties and compare it with three of the previously studied approaches: the distance-based Intra-List Diversity of our unified framework, the Intent-Aware metrics and the Proportionality framework.

4.8 re-ranking strategies

So far we have shown how our framework is able to describe different perspectives on the evaluation of novelty and diversity in Recommender Systems, especially by generalizing some metrics of the state of the art and proposing new ones, and allowing them to take into account desired properties such as position and relevance-awareness. Another contribution of this framework is the definition of a generic scheme of re-ranking strategies for the optimization of our metrics by means of using the same item novelty models to re-score the recommendations provided by baseline recommendation algorithms.

Our re-ranking approach proceeds as follows. Given a recommendation algorithm whose scoring function $s : \mathcal{U} \times \mathcal{I} \to \mathbb{R}$ determines the ranking of a recommendation, we propose applying a re-scoring by means of a linear combination between this scoring function and the novelty of the items:

$$s_{nov}(i) = (1 - \lambda)\, s(u, i) + \lambda\, nov(i) \tag{4.38}$$

where λ is the parameter that controls the trade-off between the original (relevance-oriented) score and the novelty. Typically the values of the scoring function and the novelty are in different ranges or follow different distributions, making a simple linear combination biased towards the component with variance. For that reason, we apply in practice a normalization – such as z-score $\bar{x} = \frac{x-\mu}{\sigma}$ – to both the scoring function and the novelty model. An example of the effect of such normalization can be seen in Figure 4.3. In this example, a sequence x of equidistant values between 1 and 0.05 in decreasing order is linearly combined ($\lambda = 0.5$) with random values y from a random distribution with mean 6 and variance 2. As the figure shows, on one hand the non-normalized combination is completely biased towards the component with higher value range. On the other hand, the normalized combination reflects both the decreasing trend of the sequence x and the randomness of y.

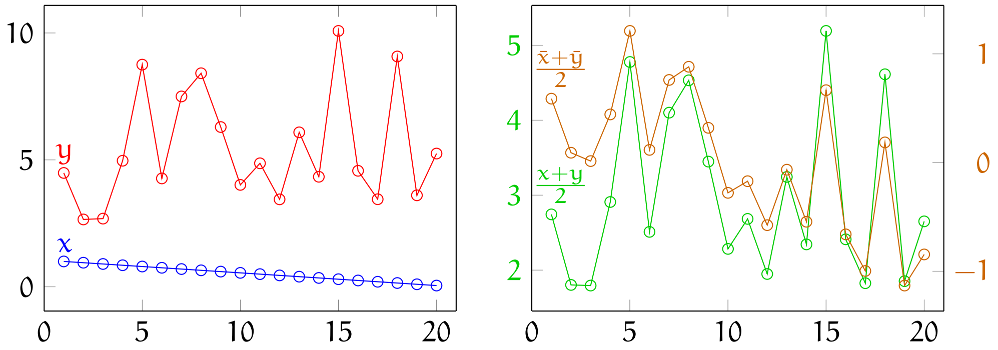

Figure 4.3: An example of linear combination (λ = 0.5) of two different value ranges with and without previous normalization.

This re-ranking scheme, unsurprisingly, generalizes or shares similarities with several proposals found in the state of the art. For instance, the work of Adomavicius and Kwon (2012) proposes a comparable re-ranking scheme in which one of its instantiations is equivalent to our re-ranking scheme with the Popularity Complement novelty model (Equation 4.2). Zhang et al. (2012) also propose a combination of a relevance score with novelty models for optimizing the serendipity of recommendations.

So far, the presented re-ranking strategy assumes that the novelty model for a recommended item can be computed before generating the final recommendation list. This is actually not the case of Intra-List Diversity and Sales Diversity, where the novelty context includes the very recommendations being generated. In those cases, our simple re-scoring approach is not valid since the novelty models we are optimizing are not completely defined until the recommendations are computed. We need thus an alternative re-ranking approach for optimizing such diversity perspectives.

A common solution found in the state of the art for optimizing Intra-List Diversity consists in performing a greedy selection by picking iteratively from the original recommendation list the items that maximize a re-scoring function with the novelty defined with respect to the set S of items already picked:

$$s_{nov}(i\,|\,S) = (1-\lambda)\; s(u,i) + \lambda\; nov(i\,|\,S) \tag{4.39}$$

Algorithm 4.1 describes this greedy selection. As commented, this greedy approach is found in previous optimization proposals for Intra-List Diversity optimization in Information Retrieval and Recommender Systems. In particular, Carbonell and Goldstein (1998) and Ziegler et al. (2005) – one for search and the other in recommendation – propose greedy selection approaches using the minimum distance of 4.1 as the novelty model of their greedy selection solutions. In the case of the Intent-Aware approaches described in Section 4.7, Santos et al. (2010a) presented

Algorithm 4.1 A greedy selection of the items in recommendation list R to produce a re-ranked list S.

```
$S \leftarrow \emptyset$
while $|R \setminus S| > 0$ do
    $i^* \leftarrow \arg\max_{i \in R \setminus S} \; s_{nov}(i \,|\, S)$
    $S \leftarrow S \cup \{i^*\}$
end while
return $S$
```

the xQuAD algorithm (see Section 2.4.2.2) which re-ranks search result lists based on the following document novelty model:

$$nov_{xQuAD}(d \,|\, S) = \sum_{s} p(s \,|\, q) \; p(d \,|\, s) \prod_{d' \in S} \left(1 - p(d' \,|\, s)\right) \tag{4.40}$$

In the case of Sales Diversity, a greedy re-ranking approach could possibly be applied. However, since the novelty context in this case is composed by all the recommendations issued to all the users, this approach turns out impractical: it would require generating all recommendations in a batch, which does not fit many scenarios where recommendations must be computed dynamically. Taking advantage of the connection between the Sales Diversity and the Long Tail Novelty approaches – further justified in our experiments in Section 4.9 –, Adomavicius and Kwon (2012) used, among others, Long Tail Novelty-based re-ranking techniques to optimize Sales Diversity. We find however this approach sub-optimal, in the sense that the perspective optimized is not the one evaluated. In Chapter 7 we propose an alternative solution in which we optimize directly Sales Diversity by (conceptually) recommending users to items.

4.9 EXPERIMENTS

In order to show the properties of our unified framework, we carry out a comprehensive set of experiments aiming to illustrate the properties of our proposed metrics and re-ranking techniques in the context of our experimental design described in Chapter 3. In particular, this evaluation aims to answer the following questions:

- How do the standard collaborative filtering algorithms behave in terms of the different metrics derived from our framework?
- What are the effects of introducing rank and relevance-awareness in the metrics?
- Are re-ranking techniques able to achieve improvements with respect to the original recommendation baselines?

- How are the different metrics related between each other? In particular, do metrics measuring the same perspective always agree? Which metrics modeling different perspectives are related?

Taking our common experimental framework as a basis, we evaluate in Section 4.9.1 some recommendation baselines with the proposed metrics of our framework. In order to do that, we compute the item novelty models that compose our metrics taking the following considerations:

- Long Tail Novelty item models (PC, FD) were calculated using our proposed estimations in Equations 4.11 and 4.12 using the training data.

- We measured Unexpectedness by using the subset of the user profiles in the training data. Two distance-based novelty models were considered: one using the collaborative filtering similarity from Equation 4.22 (PD_{CF}) and the other using the content-based similarity from Equation 4.21 ($\text{PD}_{\mathcal{F}}$) by using the genre information of each dataset.

- For measuring Intra-List Diversity we also considered the same two choices for distance as in Unexpectedness, resulting in two different distance-based novelty models (ILD_{CF} and $\text{ILD}_{\mathcal{F}}$) which differ from Unexpectedness in the fact that for ILD_{CF} we use all the data from both the training and test subsets.

- Temporal Novelty was only tested for the Netflix dataset, given the uneven temporal distribution of ratings in MovieLens1M and the absence of timestamps in the Million Song Dataset. For the Netflix data, the last month of the training data (December 2005) was removed to created one "previous" recommendation which defines a novelty model (TN) as proposed in Equation 4.14.

- The different Sales Diversity models estimated with Equation 4.18 (IUDC, IURD, IUFD) are defined by considering all the recommendations issued to the users with test data.

- Our proposed Sales Novelty novelty model (ISDC) is computed as in Equation 4.20 with respect to all the recommendations produced by the algorithms of our experimental design: random (Random), popularity (Pop), implicit Matrix Factorization (iMF), probabilistic Latent Semantic Analysis (pLSA), and user (UB) and item-based (IB) Nearest Neighbors.

To analyze the effects of rank and relevance in novelty and browsing models, we tested different combinations of rank discounts and relevance models for the metrics. Specifically, we considered two ranking models: a neutral discount (no rank) $\text{disc}(k) = 1$ that assumes all the items in the recommendation are browsed, and a exponential discount (rank) $\text{disc}(k) = \beta^{k-1}$ with $\beta = 0.9$. As for the relevance models, two alternatives were chosen: no relevance (no rel) $p(\text{rel}|i,u) = 1$ and the proposed model in Equation 4.25 with $b = 0.2$ (rel), which roughly assumes that 20% of the recommended items which are not in the test data are relevant. We thus get, for each proposed item novelty, four different resulting metrics. All these metrics were evaluated with cut-off 50, that is, $\forall u \ |R_u| = 50$. This relatively large list size was used to properly illustrate the properties of the rank-aware versions of our metrics.

We also applied the re-ranking strategies in Section 4.8 to the personalized algorithms (iMF, pLSA, UB, IB) in the MovieLens1M dataset. For each recommender, the top 100 recommended items for each user were re-ranked.

Finally, in order to examine the connection between novelty and diversity perspectives and metrics, we examine more in depth the interplay between re-ranking techniques and metrics of selected combinations of metrics from the same or different novelty and diversity perspectives in MovieLens1M. The results of cross-novelty comparisons provide further insights and discussion about the differences and connections between metrics unfolded by our framework.

4.9.1 *Evaluation of Baseline Recommendation Algorithms*

We begin our analysis by analyzing the performance with respect to our framework of some well-known collaborative filtering algorithms and the trivial but illustrative random and popularity-based recommendations. Tables 4.2, 4.3, 4.4 and 4.5 show the results for the three datasets divided by user and business-oriented metrics and relevance models. When comparing different recommendation baselines to each other, rank-awareness seemed not to result in any noteworthy effect, therefore we omit these results. Additionally, Table 4.6 includes the results of Temporal Novelty for the Netflix dataset.

Table 4.2 shows the results of user-oriented metrics with uniform relevance (no rel). As expected in such setting, random recommendations are a natural source of novelty and diversity, achieving the best results in all three datasets for all metrics. On the contrary, recommending the most popular items results in obviously poor recommendations in terms of Long Tail Novelty, with respect to which popularity-rank is by definition the worst approach. Popular recommendations also present, in general, low Unexpectedness and Intra-List Diversity when these are measured with a collaborative filtering-based distance. An exception to the previous observation is the high EPD_{CF} in the Million Song Dataset, probably caused by the fact that music is a more "niche" domain were users have more heterogeneous tastes than in the movie recommendation domain, and therefore random users may not share the tastes of the users who listened to the most popular songs. In terms of genre-based Unexpectedness and Diversity, the popularity recommender shows

		EPC	EFD	EPD$_{\text{CF}}$	EPD$_{\mathcal{F}}$	EILD$_{\text{CF}}$	EILD$_{\mathcal{F}}$
ML1M	**Rnd**	0.9668	13.4658	0.8941	0.7349	0.9157	0.7295
	Pop	0.7382	9.0189	0.7351	0.7562	0.5116	0.7132
	iMF	0.8471	10.0130	0.7527	0.6692	0.6521	0.6608
	pLSA	0.8413	9.9581	0.7494	0.6719	0.6358	0.6540
	UB	0.7958	9.5036	0.7337	0.6925	0.5768	0.6876
	IB	0.8097	9.6731	0.7369	0.6950	0.5975	0.6981
Netflix	**Rnd**	0.9866	16.2202	0.9383	0.7709	0.9533	0.7885
	Pop	0.7496	9.4016	0.7318	0.7446	0.4841	0.7165
	iMF	0.8533	10.5020	0.7508	0.6966	0.6253	0.6767
	pLSA	0.8303	10.1478	0.7391	0.7007	0.5832	0.6840
	UB	0.8141	9.9698	0.7355	0.7093	0.5612	0.7034
	IB	0.8122	10.0250	0.7292	0.7127	0.5520	0.7169
MSD	**Rnd**	0.9998	20.5046	0.9991	0.8859	0.9997	0.9072
	Pop	0.9688	10.4738	0.9717	0.8092	0.8455	0.7340
	UB	0.9854	12.9169	0.9539	0.7201	0.9243	0.7011
	IB	0.9937	15.4715	0.9472	0.6985	0.9304	0.6936

Table 4.2: Results of the user-oriented metrics of the unified framework without rank or relevance for the baseline recommendation algorithms in MovieLens1M, Netflix and Million Song Dataset.

high scores for both, the first probably caused by the non-personalized nature of the algorithm and the second by the inherent variety in the collective preferences of the users.

Regarding the personalized recommendation algorithms, we see particular outcomes for each dataset. In the case of MovieLens1M, we see a clear difference between latent factors and nearest neighbors algorithms. On one hand, latent factors algorithms produce totally different results to the most-popular recommendation: highly novel and diverse recommendations in terms of EPC, EFD, EPD$_{\text{CF}}$ and EILD$_{\text{CF}}$ and low scores in term of the genre-based metrics. On the other hand, nearest neighbors are a middle point between popularity and latent factors algorithms in all metrics but EPD$_{\text{CF}}$, where they perform as bad as popularity. Such behavior is expected, especially for the item-based variant, which is based on maximizing the similarity of the recommended items to those of the user's profile. In the case of the Netflix, the results are highly similar, but in this case pLSA is not as competitive as iMF. In the Million Song Dataset the nearest neighbors algorithms present opposite properties to the popularity recommendation.

		EPC	EFD	EPD$_{CF}$	EPD$_{\mathcal{F}}$	EILD$_{CF}$	EILD$_{\mathcal{F}}$
ML1M	**Rnd**	0.1966	2.7316	0.1817	0.1495	0.1862	0.1486
	Pop	0.1931	2.3628	0.1920	0.1952	0.1342	0.1883
	iMF	0.2659	3.1486	0.2378	0.2122	0.2026	0.2083
	pLSA	0.2539	3.0121	0.2281	0.2052	0.1902	0.1995
	UB	0.2381	2.8446	0.2199	0.2063	0.1715	0.2045
	IB	0.2347	2.8066	0.2145	0.2013	0.1722	0.2017
Netlfix	**Rnd**	0.1985	3.2595	0.1887	0.1552	0.1919	0.1588
	Pop	0.1884	2.3606	0.1828	0.1858	0.1216	0.1805
	iMF	0.2540	3.1260	0.2238	0.2086	0.1850	0.2023
	pLSA	0.2433	2.9742	0.2168	0.2063	0.1693	0.2012
	UB	0.2399	2.9364	0.2164	0.2090	0.1636	0.2068
	IB	0.2304	2.8415	0.2068	0.2021	0.1559	0.2028
MSD	**Rnd**	0.2000	4.1016	0.1999	0.1772	0.2000	0.1815
	Pop	0.2036	2.2001	0.2038	0.1695	0.1777	0.1546
	UB	0.2403	3.1936	0.2304	0.1717	0.2236	0.1665
	IB	0.2502	3.8624	0.2360	0.1721	0.2325	0.1710

Table 4.3: Results of the user-oriented metrics of the unified framework without rank but considering relevance ($b = 0.2$) for the baseline recommendation algorithms in MovieLens1M, Netflix and Million Song Dataset.

When we consider the relevance for assessing our user-oriented novelty and diversity metrics, as Table 4.3 shows, we get different and insightful results. In this new setting – which assigns a background relevance probability of $b = 0.2$ for items absent in the test data – the random recommendations are heavily penalized by their lack of relevance. Only in a few cases, such as EFD, the overwhelmingly high novelty of the random recommendations compensates for their lack of relevance and results in competitive values. The most-popular recommendations, when considering relevance, result in poor scores of novelty and diversity: their strengths in the relevance-unaware case are now weakened because of their relatively low relevance compared to their personalized counterparts. In the case of the personalized recommendations, we see that introducing relevance-awareness has the desirable effect of balancing the values of novelty and diversity with the relevance of the algorithms. The cases of movie recommendations have some illustrative examples: latent factor models had low genre-based Unexpectedness and Intra-List diversity compared to nearest neighbors methods, but including the high relevance of the former in the metrics makes them comparable with the latter. For

		EIUDC	EIURD	EIUFD	EISDC
ML1M	Rnd	0.9862	73.6040	11.8340	0.9859
	Pop	0.3803	4.5840	6.5361	0.7109
	iMF	0.9034	33.6920	9.5212	0.6330
	pLSA	0.8944	32.8200	9.4476	0.6192
	UB	0.7787	22.6960	8.4763	0.5746
	IB	0.7798	34.4245	8.5297	0.6282
Netflix	Rnd	0.9946	186.4000	13.1778	0.9946
	Pop	0.3786	5.9000	6.5572	0.7368
	iMF	0.9050	91.5160	9.7213	0.6844
	pLSA	0.8720	46.3680	9.1344	0.6327
	UB	0.8324	111.4029	8.7781	0.6209
	IB	0.7743	192.7314	8.5553	0.6484
MSD	Rnd	0.9997	3,016.3600	17.1732	0.9997
	Pop	0.0212	1.4600	5.6941	0.8718
	UB	0.8798	1,357.1200	11.5035	0.7885
	IB	0.9789	2,873.9049	14.2542	0.8379

Table 4.4: Results of the business-oriented metrics of unified framework without rank and relevance for the baseline recommendation algorithms in MovieLens1M, Netflix and Million Song Dataset.

example, Figure 4.4 shows the change in $\text{EPD}_{\mathcal{F}}$ caused by considering relevance in MovieLens1M.

The results of the relevance-unaware business-oriented metrics can be found in Table 4.4. Again, random recommendations offer the best results in terms of Sales Diversity and Novelty when relevance is not considered. Popular recommendations, as expected, have by far the lowest values in Sales Diversity metrics, although its Sales Novelty measured by EISDC is higher than the personalized algorithms. The reason for this is clear: when measuring the novelty of sales with respect to other four personalized algorithms, it is expected that these will be more similar between each other, and therefore a non-personalized algorithm, however obvious as popularity, is expected to be different to the personalized ones. Regarding the personalized algorithms, each dataset has its own particularities. In MovieLens1M, latent factors algorithms have the highest Sales Diversity values, with the exception of EIURD, in which the item-based nearest neighbors presents the highest score. In terms of Sales Novelty, the user-based is clearly the worst option, while the rest of personalized algorithms perform similarly. A similar outcome is observed in

		EIUDC	EIURD	EIUFD	EISDC
ML1M	Rnd	0.2009	15.0278	2.4111	0.1995
	Pop	0.1029	1.6020	1.7320	0.1657
	iMF	0.2844	10.3131	2.9893	0.1873
	pLSA	0.2708	8.9244	2.8450	0.1728
	UB	0.2327	6.5997	2.5328	0.1580
	IB	0.2266	8.9075	2.4756	0.1648
Netlfix	Rnd	0.2003	37.5718	2.6541	0.1998
	Pop	0.0999	2.2435	1.6719	0.1675
	iMF	0.2697	28.9170	2.9002	0.1927
	pLSA	0.2562	15.2927	2.6855	0.1726
	UB	0.2459	27.9245	2.5957	0.1699
	IB	0.2209	42.3997	2.4293	0.1690
MSD	Rnd	0.2000	603.4007	3.4354	0.2000
	Pop	0.0046	0.3189	1.1979	0.1801
	UB	0.2161	434.0217	2.8789	0.1873
	IB	0.2465	703.5852	3.5814	0.2047

Table 4.5: Results of the business-oriented metrics of the unified framework without rank but considering relevance ($b = 0.2$) for the baseline recommendation algorithms in MovieLens1M, Netflix and Million Song Dataset.

the Netflix data, where according to EIUDC and EIUFD, the best to worst order is defined as iMF, pLSA, UB and IB, but EIURD gives an almost opposite order: IB, UB, iMF and pLSA. In terms of EISDC, the implicit matrix factorization is in the Netflix data the most novel choice, while the user-based algorithm performs slightly worse than the rest. Finally, in the Million Song dataset, the item-based algorithm is in all metrics better than the user-based variant.

Now, adding relevance to Sales Diversity and Novelty metrics produces the outcomes shown in Table 4.5. The random recommendation, despite its low relevance, achieves competitive results in terms of EIURD, EIUFD and EISDC given its near-to-perfect high values in the relevance-unaware setting and the help of the background relevance probability. In the case of EIUDC, which gives a much lower reward for the highly novel items than the other Sales Diversity alternatives, the random recommendation scores below the personalized recommendations. The relative performance of the most-popular recommendations is unchanged with respect to Sales Diversity. Moreover, its high value in Sales Novelty for the relevance-unaware case is not sufficient to compensate for its low relevance, thus performing

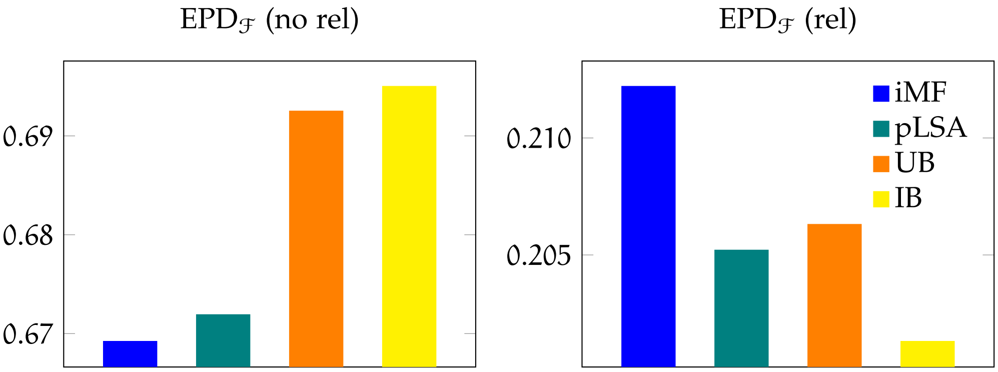

Figure 4.4: Genre-based EPD without and with relevance in MovieLens1M.

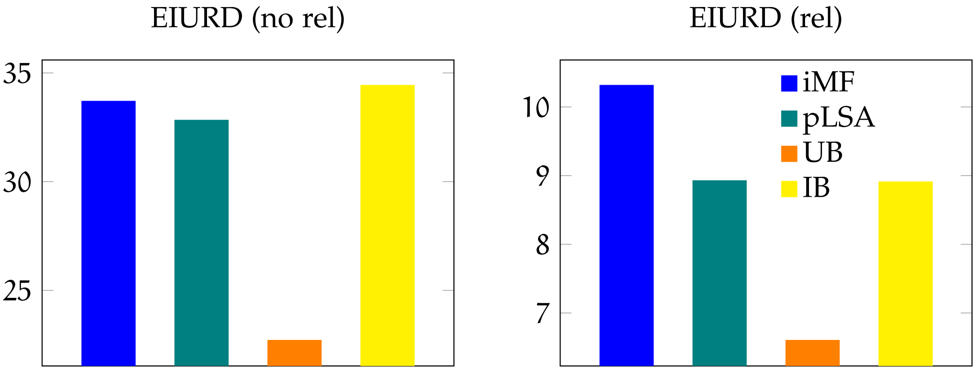

Figure 4.5: EIURD without and with relevance in MovieLens1M.

poorly compared to personalized algorithms. In the case of personalized recommendations, no significant changes in Sales Novelty are observed when considering the relevance of the recommendations. However, introducing relevance does evidence some changes in Sales Diversity metrics for MovieLens1M and Netflix. In the first dataset, the high performance of the IB in EIURD is moderated by its lower relevance compared with the latent factors algorithms, as shown in detail in Figure 4.5. In the second dataset, the results of EIURD are also modified by the relevance, in particular now iMF scores higher than UB due to their similar values without relevance but higher relevance of iMF.

The results for Temporal Novelty in the Netflix dataset are shown in Table 4.6. As expected, random recommendations show the best values in terms of the ETD metric, even when considering relevance. In turn, the most-popular recommendations barely change from one month to another, resulting in the lowest values of temporal change between recommendations. Regarding the more elaborate personalized recommendations, here pLSA is the best choice, since it is capable of recommending around 40% of new movies over time – when not considering relevance. The item-based algorithm is, in turn, the most static recommendation after one month of activity. The results for the personalized algorithms do not vary in relative terms when relevance is considered, due to the large differences in terms of uniform rel-

	ETD (no rel)	ETD (rel)
Rnd	0.9945	0.2009
Pop	0.0302	0.0076
iMF	0.2140	0.0595
pLSA	0.3839	0.1047
UB	0.1445	0.0380
IB	0.0972	0.0252

Table 4.6: Results of rank-unaware Temporal Diversity in temporal partition of Netflix.

evance novelty. Interestingly, these results contrast with the observations of Lathia et al. (2010), who observed that item-based nearest neighbors outperform matrix factorization in terms of temporal novelty. We think this discrepancy is caused by the differences in the choice of the variants of the baseline algorithms and the evaluation methodology: in the work of (Lathia et al., 2010) their recommendations target rating prediction and their accuracy is therefore measured by prediction error, while our algorithms and evaluation aims at maximizing the prediction of ranked lists.

In short, the evaluation of the baselines recommendations offers the following conclusions:

- Random recommendations are a natural source of novel and diverse recommendations when relevance is not considered, while most-popular recommendations clearly offer generally bad results for certain perspectives on novelty and diversity of recommendations.
- Where the comparison is possible, latent factors offer more novel results than nearest neighbors algorithms, although the latter offer more variety within recommendations than the former as measured by the features of the items.
- In this experiments, rank-awareness clearly affects the comparison between baselines – notably for the non-personalized algorithms – by balancing novelty and diversity with relevance of results, while rank-awareness does not offer any noteworthy effect in this setting.

4.9.2 *Evaluation of Re-Ranking Strategies*

The results of applying re-ranking techniques to the personalized recommendations in the MovieLens dataset are displayed in Figure 4.6. For each recommendation baseline, we applied a re-ranking of the top-100 recommendations with the normalized linear combination between the original scores provided by each recommendation algorithm and the values of the novelty model that is measured. We show different values of the λ parameter that controls the linear combination starting from $\lambda = 0.0$ (no re-ranking) to $\lambda = 1.0$ (top-100 sorted by novelty) by steps of 0.1. We measured the performance of these re-rankings with three variants of each target metric: one without ranking or relevance discount (no rank, no rel), a second one that only takes relevance into account (no rank, rel) and a third one with both rank and relevance components. For brevity, the results for the variant with ranking but not relevance (rank, no rel) are not included since they do not differ significantly from the first (no rank, no rel) alternative.

The plots in the **first column** of Figure 4.6 show the effect of re-ranking strategies for the rank and relevance-unaware (no rank, no rel) novelty and diversity metrics of our framework. As expected, the higher the weight of the novelty component in the re-ranking, the better are the results in the target metric. An interesting observation in these results is that, for each value of λ, the novelty and diversity values of the re-ranked recommendations maintain the relative order obtained by the original baseline ($\lambda = 0.0$).

The **second column** shows the results when considering the relevance of the re-ranked items (no rank, rel). It is expected that modifying the original order of the recommendation will result in a loss in terms in relevance and, therefore, we expect to see a trade-off between the improved novelty of the re-rankings and their expected loss in relevance. As foreseen, this relevance-aware evaluation of the re-ranking strategies show, in general, that applying an excessive weight on the novelty component results in a loss in relevance-aware novelty and diversity metrics. Instead, more moderated trade-offs in the re-ranking strategies tend to offer improvements over the results of the baselines with varying success for each metric and baselines. For instance, in Long Tail Novelty metrics (EPC and EFD) the latent factors algorithms do not show notable improvements, probably due to their high values in these metrics without re-ranking steps, and only nearest neighbors solutions seem to show some perceivable improvements. In the case of EPD_{CF}, only the re-ranking of the UB show some improvements with respect to the baseline. For the rest of the user-oriented metrics, applying an intermediate re-ranking ($\lambda = 0.5$) seems to offer the best results. Finally, the results of EISDC do not seem to be specially affected by a loss in the relevance of the re-ranked recommendations.

The **third column** shows the results when considering both rank and relevance (rank, rel) in the metrics. The results show some interesting changes with respect to the second column (no rank, rel). In particular, adding rank-awareness seems to further penalize the results of full ($\lambda = 1.0$) novelty-oriented re-rankings for all metrics. An illustrative example of the former observation is the rank-aware EISDC, that shows a drop when applying a too aggressive re-ranking. Another interesting observation is that, according to this rank and relevance combination,

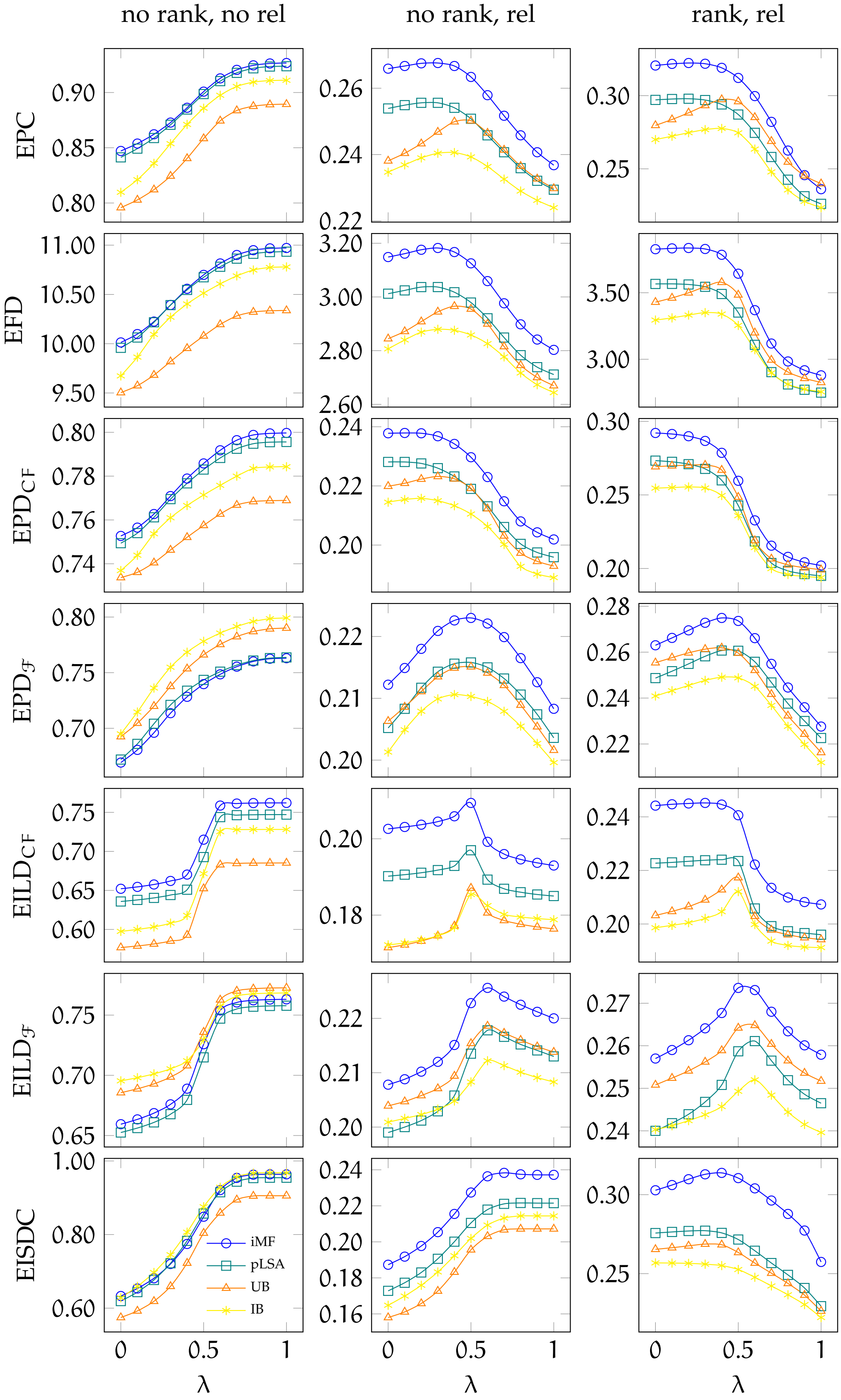

Figure 4.6: Re-ranking techniques applied to their target metrics of the unified framework in the MovieLens1M dataset.

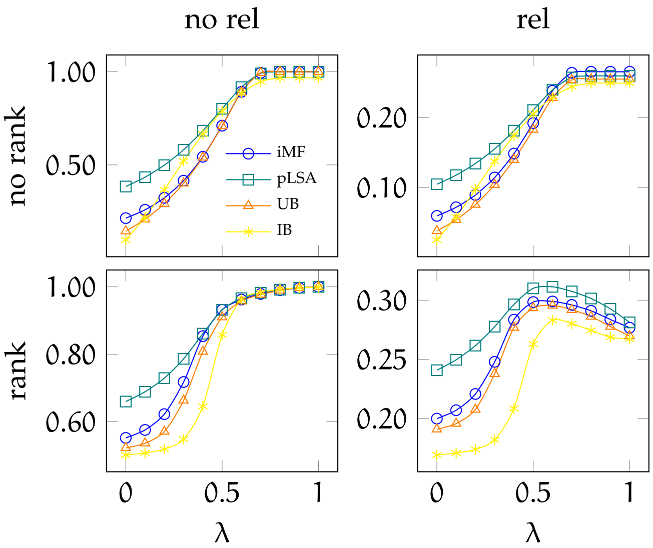

Figure 4.7: Re-ranking techniques and metrics for Temporal Novelty in Netflix.

the re-ranking of the UB algorithm is now competitive when compared to pLSA, especially in EPC and EFD.

Finally, we also applied a re-ranking of the results of Temporal Novelty for the Netflix dataset. The results are shown in Figure 4.7. In this case rank-awareness is the property that establishes a bigger difference with respect to the original (no rank, no rel) metric. Applying a temporal novelty-oriented re-ranking seems to offer noticeable improvements with respect to the baseline. The case of IB is here specially interesting: according to the original (no rank, no rel) metric, the re-ranking technique is able to make this algorithm competitive with respect to the rest. However, in the context of rank-awareness, such relative improvements of the IB algorithm are mitigated.

4.9.3 *Relation between Novelty and Diversity Metrics*

We turn now to analyzing the relationships between the different notions of novelty and diversity and the connections between different metrics arising from our framework and other alternatives. Our discussion will be based on the results of baseline recommendations for all datasets and, particularly, in the re-rankings applied to the user-based algorithm – the one that seems to benefit the most from re-ranking across metrics – in the MovieLens1M dataset. In particular, re-ranking provides a natural way of relating different metrics and perspectives: if applying a re-ranking targeting one metric achieves consistent improvements with respect to another, we get evidence for the degree of relatedness of both metrics.

		EPC	EFD	EIUDC	EIURD	EIUFD	EISDC
no rel	**UB**	0.7957	9.5033	0.7787	22.6960	8.4763	0.5746
	+**PC** (1.0)	0.8894	10.3366	0.9009	29.2680	9.3831	0.7990
	+**FD** (1.0)	0.8894	10.3366	0.9009	29.2680	9.3831	0.7990
	+**ISDC** (1.0)	0.8486	9.9454	0.8857	28.9720	9.2066	0.9054
rel	**UB**	0.2380	2.8445	0.2327	6.5997	2.5328	0.1580
	+**PC** (0.5)	0.2502	2.9412	0.2666	8.0349	2.7630	0.1787
	+**FD** (0.4)	0.2501	2.9653	0.2653	8.0890	2.7758	0.1720
	+**ISDC** (0.8)	0.1957	2.2915	0.2040	6.6848	2.1206	0.2072

Table 4.7: Discovery-based re-rankings and metrics for the user-based nearest neighbors algorithm in MovieLens1M.

4.9.3.1 *Relation between Discovery-Based Metrics*

In Section 4.9.1 we observed how latent factors were particularly good in terms of Long Tail Novelty, Sales Diversity and Sales Novelty, whilst nearest neighbors offered low performance in all three perspectives – with some particular exceptions, such as the results of EIURD. We are therefore interested in further analyzing this connection between all these discovery-based metrics which measure different notions of novelty and discovery. In Table 4.7 we explore the interplay of applying discovery-based re-ranking techniques and their effect on other non-targeted discovery-based metrics. The results show additional insights that confirm our suspected connection between these perspectives. In particular, we see that, when the metrics do not take into account the relevance of the recommendations, all re-rankings of the UB algorithm improve all discovery-based metrics. When considering the relevance, we see that the Sales Novelty-oriented re-ranking degrades the performance for the other considered metrics, but Long Tail Novelty re-rankings still achieve enhancements for all Sales Diversity metrics. This observation matches the findings of Adomavicius and Kwon (2012).

We can conclude that, according to our discovery-based metrics, Long Tail Novelty, Sales Diversity and Sales Novelty are naturally related. The reason for this connection lies in the popularity bias of most common recommendations scenarios 2.2.4. As we observed in Section 3.2, user-item interaction usually displays a long tail effect among items. Collaborative filtering algorithms make use of these user-item interactions to create recommendations, so it is expected that these Long Tail effects will also show up in the resulting recommendations, therefore affecting the results of Sales Diversity and Sales Novelty.

		EPD$_{\text{CF}}$	**EPD**$_{\mathcal{F}}$	**EILD**$_{\text{CF}}$	**EILD**$_{\mathcal{F}}$
no rel	**UB**	0.7337	0.6925	0.5768	0.6898
	+**PD**$_{\text{CF}}$ (1.0)	0.7689	0.6917	0.6807	0.6904
	+**PD**$_{\mathcal{F}}$ (1.0)	0.7461	0.7900	0.6002	0.7441
	+**ILD**$_{\text{CF}}$ (1.0)	0.7665	0.6850	0.6852	0.6898
	+**ILD**$_{\mathcal{F}}$ (1.0)	0.7431	0.7416	0.6054	0.7768
rel	**UB**	0.2199	0.2063	0.1715	0.2052
	+**PD**$_{\text{CF}}$ (0.3)	0.2232	0.2057	0.1810	0.2042
	+**PD**$_{\mathcal{F}}$ (0.5)	0.2106	0.2151	0.1632	0.2144
	+**ILD**$_{\text{CF}}$ (0.5)	0.2204	0.1989	0.1871	0.1981
	+**ILD**$_{\mathcal{F}}$ (0.6)	0.2147	0.2112	0.1698	0.2199

Table 4.8: Distance-based re-rankings and metrics for the user-based nearest neighbors algorithms in MovieLens1M.

4.9.3.2 *Relation between Distance-Based Metrics*

In the results for Unexpectedness and Intra-List Diversity in Section 4.9.1 we observed that the two different distance models between items – content-based and collaborative filtering-based – resulted in very different outcomes for the baseline algorithms in the three datasets. As suggested in Section 4.5.2, there is little evidence that both approaches are correlated. Moreover, we also saw that, when using the same distance measurement for EPD and EILD, algorithms that were good in one also performed adequately in the other, and vice versa.

To confirm our preliminary observations from the baselines results, we tested the re-ranking techniques and metrics of the four distance-based metrics – EPD$_{\text{CF}}$, EPD$_{\mathcal{F}}$, EILD$_{\text{CF}}$ and EILD$_{\mathcal{F}}$ – with each other. The results in Table 4.8 confirm two observations: collaborative filtering-based distance is quite different from content-based distance and, under the same metric, improvements in Unexpectedness result in increased Intra-List Diversity and vice versa. There is however an interesting comparison between distance measurements without relevance: the genre-based re-rankings actually brings improvements in terms of collaborative filtering-based metrics, but not the opposite. This finding provides a weak but perceivable confirmation of our intuition expressed in Section 4.5.2: given two items with many common users, the chances of having similar content is higher than for two random items without common users. When we consider relevance in the assessments though, the re-rankings with one distance hurt the performance of the other in both Unexpectedness and Intra-List Diversity. Finally, the relation between equally-measured EPD and EILD can be explained by the personalized nature of the baseline algorithm. We expect personalized recommendation algorithms to provide

	$\text{EILD}_{\mathcal{F}}$ (no rel)	$\text{EILD}_{\mathcal{F}}$ (rel)	ERR-IA	α-nDCG	S-recall
UB	0.6898	0.2052	0.2085	0.4066	0.8314
+$\text{ILD}_{\mathcal{F}}$ (0.6)	0.7669	0.2199	0.1930	0.3750	0.8702
+**xQuAD** (0.4)	0.6870	0.2048	0.2292	0.4403	0.8547

Table 4.9: Distance and Intent-Aware re-rankings and metrics for measuring Intra-List Diversity for the user-based nearest neighbors algorithms in MovieLens1M.

items similar to the tastes of the user in her profile, and therefore, similar to each other.

4.9.3.3 *Relation to Alternative Intra-List Diversity Metrics*

Finally, we consider the relation between our approach for measuring Intra-List Diversity, the EILD metric, and the alternative metrics outside of our framework introduced in Section 4.7: the Intent-Aware metrics ERR-IA and α-nDCG (with $\alpha = 0.5$) and S-recall. Although these metrics are defined for search result diversification, we make a preliminary translation by substituting queries by users, documents by items and subtopics by feature information – in our case genres. In Chapter 5 we will deepen into such equivalences and the applicability of the Intent-Aware Diversity in Recommender Systems.

The results in Table 4.9 reveal the differences between our proposal for measuring Intra-List Diversity and the approaches of the Intent-Aware framework. As we can see, improvements in terms of $\text{EILD}_{\mathcal{F}}$ are not translated to the Intent-Aware metrics, and the xQuAD re-ranking algorithm for enhancing Intent-Aware metrics seems to effectively optimize its target metrics, but not $\text{EILD}_{\mathcal{F}}$. Interestingly, both re-ranking techniques are able to improve the total number of genres retrieved as measured by S-recall.

The evidence points to a divergence between the criteria of both approaches for measuring and enhancing Intra-List Diversity. Despite using similar feature-based information, these two perspectives take different approaches in their novelty models, and they only seem to agree in improving the recall of subtopics or genres.

4.10 CONCLUSIONS

The research presented here aims to contribute to a shared characterization and understanding of the basic elements involved in recommendation novelty and diversity upon a formal foundation. The proposed framework provides a common ground for the development of metrics based on different perspectives on novelty and diversity, generalizing metrics reported in the literature, and deriving new ones. An advantage of the proposed decomposition into a few essential modular pieces is a high potential for generalization and unification. Two novel features in novelty and diversity measurement arise from our study: rank sensitivity, and relevance awareness. Both aspects are introduced in a generalized way by easy to configure components in any metric supported by our scheme. Our experiments validate the proposed approach and provide further observations on the behavior of metric variants.

5 INTENT-AWARE DIVERSITY IN RECOMMENDER SYSTEMS

5.1 INTRODUCTION

Search result diversification is being actively researched in the Information Retrieval community as a means to address query ambiguity and underspecification in ad-hoc Information Retrieval tasks (Carbonell and Goldstein, 1998; Agrawal et al., 2009; Clarke et al., 2008). In general terms, and most particularly in common practical scenarios, recommendation can be seen as an Information Retrieval task. Under this vision it would seem natural to consider a connection between diversity as researched in the IR field and notions of diversity and novelty developed in Recommender Systems. However, the diversity issue has been stated and addressed rather differently in the research on the topic so far in Recommender Systems and Information Retrieval respectively. On one hand, diversity has been studied under a quite specific motivation and precise problem definition in the Information Retrieval community – building around the problem of uncertainty in user queries – along with formally grounded and well understood diversity metrics, with a significant theoretical development and a drive towards standardization (backed by a specific TREC diversity task (Clarke et al., 2009, 2010, 2011b, 2012)). On the other hand, as pointed out in Chapter 4, such level of formalization and standardization regarding novelty and diversity is rather missing in the area of Recommender Systems. It seems therefore natural to wonder whether, as far as it were possible to draw models and principles from one area to the other, research on Recommender Systems diversity might benefit from the insights and ongoing progress in search result diversification.

In this chapter we explore the adaptation of diversity models, metrics, and methods from ad-hoc Information Retrieval, specifically what we denote as **Intent-Aware framework**, which comprises all those approaches that use explicit aspects to diversify search results (Clarke et al., 2008; Agrawal et al., 2009; Santos et al., 2010a), into a Recommender Systems setting. We propose the notion of **user aspect** as an analogue of query intent, upon which we adapt the Information Retrieval diversity techniques and methodology to the recommendation task. In particular, we consider the case of aspect spaces defined by the features of the items in the user profiles, making a suitable choice of the item feature space for the purpose of defining and assessing diversity.

On top of this adaptation, we propose two new methods to enhance the diversity of recommendations. The first method takes an **explicit relevance model** for explicit diversification techniques. Intent-Aware search result diversification methods typically rely on generative views of the retrieval system to be diversified, assuming implicit relevance (Agrawal et al., 2009; Santos et al., 2010a). Our approach provides an alternative understanding of this problem by explicitly including a formal relevance model, which entails a new theoretical perspective on the problem. This approach shows a competitive or better performance than its generative-based counterparts and allows further extensions, in particular for a graded redundancy management.

The second method uses **user sub-profiles** to exploit the diversity of user preferences for our adaptation of Intent-Aware diversification methods. A particularly effective approach for the extraction of query intents uses query reformulations returned by a search engine as proxies for query aspects (Santos et al., 2010a). Drawing from this perspective, we propose the adaptation of the notion of query reformulation for recommendations through the extraction of user sub-profiles. Considering subsets of user interests is a natural idea, since people's preferences have different sides (sports, politics, work, leisure, music or movie genres, etc.), as well as we have different facets in our lives, and different attitudes in different contexts. The definition of user sub-profiles seeks to make specific recommendations to a user according to every single facet or interest. The basis of the approach we propose in this chapter is the intuition that better (more accurate and better diversified) recommendations can be produced by taking into account this multifaceted nature of user interests. The idea is that, for instance, user preferences in classical music can be more useful than rock music favorites to recommend a classical music piece. Our approach identifies the diversity within user profiles and generates partial recommendations based on homogeneous subsets of user preferences (sub-profiles), which we combine later to produce a final recommendation.

We report experiments on the context of the experimental design described in Chapter 3. This evaluation provides, first, a confirmation of the soundness of our adaptation of Intent-Aware metrics and methods to the recommendation problem. Second, it shows the advantages of our two diversification proposals – explicit relevance models and user sub-profiles – compared to the direct adaptations of the methods in the state of the art.

The rest of the Chapter is structured as follows. In Section 5.2 we define the concept of user aspects and show how to extract them from feature information for the items in the user profiles. With such user aspects, in Section 5.3 we adapt some well-known metrics and diversification methods from the state of the art to our recommendation setting. Our proposals for considering explicit relevance models and user sub-profiles are detailed in Section 5.4 and Section 5.5, respectively. We show the validity of our adaptation of the Intent-Aware framework for

Recommender Systems and our two proposals in an experimental evaluation in Section 5.6. Finally, Section 5.7 offers the conclusions.

The contents of this chapter have been presented in following published work:

- Vargas, S. and Castells, P. (2013). Exploiting the diversity of user preferences for recommendation. In *Proceedings of the 10th Conference on Open Research Areas in Information Retrieval*, OAIR '13, pages 129–136, Paris, France. CID

- Vargas, S., Santos, R. L. T., Macdonald, C., and Ounis, I. (2013). Selecting effective expansion terms for diversity. In *Proceedings of the 10th Conference on Open Research Areas in Information Retrieval*, OAIR '13, pages 69–76, Paris, France. CID

- Vargas, S. and Castells, P. (2012). Diversificación en sistemas de recomendación a partir de sub-perfiles de usuario. In *II Congreso Español de Recuperación de Información*, CERI'12

- Vargas, S., Castells, P., and Vallet, D. (2012a). Explicit relevance models in intent-oriented information retrieval diversification. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '12, pages 75–84, New York, NY, USA. ACM

- Vargas, S., Castells, P., and Vallet, D. (2011). Intent-oriented diversity in recommender systems. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '11, pages 1211–1212, New York, NY, USA. ACM

5.2 user aspects

As stated in the introduction, our adaptation of Intent-Aware methods of Search Result Diversification is performed by establishing an analogy between query aspects or intents with a notion of user aspect. These user aspects should be suitable to capture in some way, for each user, the variety and heterogeneity of tastes and interests laying within a single user. The variety of interests and trends of an individual person introduces an element of ambiguity in matching the user's need in a recommendation, which connects with the problem of query ambiguity and underspecification in search, and therefore diversity as a means to cope with this uncertainty. In this section we elaborate on the relation of the problem of diversity in both domains. We present the notion of user aspects and provide practical means of extracting such aspects by the use of the preferences expressed in the user profile.

5.2.1 *Recommendation Diversity vs. Search Diversity*

Diversity in Recommender Systems is generally motivated as a means to deal with redundancy based on the assumption that recommending too similar items is less profitable for the user – and the business – than offering a more varied experience. Looking back for a connection to diversity in ad-hoc Information Retrieval, one finds that the issues of ambiguity and underspecification are generally absent from the problem statement in the Recommender Systems literature. This may seem natural as far as there is no query in the recommendation task to begin with. However, there is certainly a user information need, expressed in the form of a user profile (ratings or item access records). This implicit information need expression arguably involves far more ambiguity and incompleteness than an explicit user query, whereby the uncertainty-oriented motivation would certainly hold for Recommender Systems diversity. So does the principle of diversification as a means to minimize the risk of under-performance extremes, which is also common in the Information Retrieval literature (Agrawal et al., 2009).

Query ambiguity and underspecification are modeled in terms of query interpretations, categories, aspects, nuggets, subtopics, and similar elements in ad-hoc Information Retrieval. An analogy can be drawn in the Recommender Systems setting by considering an equivalent notion of **user aspect**. This is in fact a natural idea, since a single user's interests have many different sides and subareas (e.g. professional, politics, movies, travel, etc.). Different user aspects can be relevant or totally irrelevant at different times therefore, similar to query intent, there is uncertainty at recommendation time about what area of user interest should play in the given context.

By means of this analogy between query and user aspects, we provide a way of adapting theories and metrics in search diversity to the recommendation task. This adaptation brings benefits such as:

- a new perspective and rationale for diversity in Recommender Systems in terms of theory and models,
- new diversity metrics for Recommender Systems, such as the Intent-Aware metrics (Agrawal et al., 2009) or α-nDCG (Clarke et al., 2008),
- new diversification methods, such as the explicit Query Aspect Diversification of Santos et al. (2010a), and
- a step towards a shared consensus on common metrics and methodologies.

5.2.2 *Defining and Extracting User Aspects*

Formally, we consider the space of aspects $\mathcal{A}$ that represents the different and disjoint interests and tastes of the users. As these interests are not always equally representative of the user profile, we find it convenient to represent, for each user u, her aspect space as a probability distribution over the aspects that will be denoted as $p(a \mid u)$ for each aspect $a \in \mathcal{A}$. Considering, as a simplification, that these aspects cover completely and without overlap all the possible different interests of the users, we further assume that $\sum_{a \in \mathcal{A}} p(a \mid u) = 1$. This assumption is in accordance with prior work in search result diversification (Agrawal et al., 2009; Santos et al., 2010a) and, in a sense, determines that users are not be interested in more than one aspect simultaneously. This may introduce some limitations to the approach when applied to the recommendation setting, as we shall see in Chapter 6.

Throughout this chapter, we focus on the definition of aspects by means of the characteristics of the items in user profiles. In particular, given a set of features $\mathcal{F}$ of the items of our recommendation domain – such as categories, genres, tags, etc. –, we consider the aspects defined by those, that is, $\mathcal{A} = \mathcal{F}$. Under this view, the aspect distribution can be estimated as follows:

$$p(f \mid u) = \frac{|\{i \in \mathcal{I}_u \ : \ f \in \mathcal{F}_i\}|}{\sum_{f' \in \mathcal{F}} |\{i \in \mathcal{I}_u \ : \ f' \in \mathcal{F}_i\}|} \tag{5.1}$$

Alternative methods for extraction of user aspects could make use of additional information about users – such as demographic information or explicit surveys about her interests – or the latent semantic of the interactions between users and items, as in the work of Zhang et al. (2012).

The definition of user aspects by means of item features introduces some particularities that are not present in common search result diversification scenarios:

- In the dominant search diversity task formulations (to be more precise, in the TREC diversity task formulation), the set of possible subtopics of a query is inherent and unique to the query – "java download" and "java indonesia" are subtopics of the query "java", but not of the query "apple". In our feature-based aspects, a common set of features $\mathcal{F}$ represents the aspects of all the users. That is, the aspect is fixed and exists before the development of user profiles, and any user may in principle display to some degree any aspect. In this sense, our formulation resembles that of Agrawal et al. (2009), where ODP categories are shared among queries to represent their possible subtopics.

- In a search scenario, a document covering one of the intents of the query is always considered relevant. In our feature-based aspects, an item may cover an aspect of the user but not be relevant – a user may like Action movies in general, but not every action movie.

- In the traditional evaluation of search diversity – such as the diversity tasks in the TREC Web Track (Clarke et al., 2009, 2010, 2011b, 2012) – the subtopics are hidden to the Information Retrieval system and/or diversification algorithms under evaluation, and are used only for evaluation purposes. In our case, the features of the items are always available and necessary for the system to properly define the distribution of aspects for each user.
- Related to the previous point, another drawback of diversity evaluation in TREC is the lack of information about the importance of each subtopic, which results in the simplifying assumption of uniform subtopic importance for each query in the evaluation. In our case, the user profiles are a source of evidence for the estimation of the importance of each user aspect $p(a|u)$ in both diversification and evaluation.

5.3 ADAPTATION OF INTENT-AWARE METRICS AND METHODS

User aspects, as defined in the previous section, provide a natural way for adapting metrics and methods of explicit aspect-based search result diversity to recommendation. In this section, we provide examples of the adaptation of well-known metrics and diversification methods in Information Retrieval.

5.3.1 *Adaptation of Diversity Metrics*

Zhai et al. (2003) proposed the Subtopic Recall metric (S-recall, see Section 2.4.1.1). This metric computes, for a given result list, the proportion of covered subtopics of the issued query. We can easily adapt this metric by considering, for a recommendation R, the proportion of items that cover the interests represented by the aspect space:

$$\text{S-recall}(R) = \frac{|\bigcup_{i \in R} \mathcal{A}_i|}{|\mathcal{A}|} \tag{5.2}$$

where $\mathcal{A}_i$ denotes the subset of aspects covered by item i. This adapted S-recall shows some of the particularities of our adaptation:

- the possibility of considering the contribution to S-recall of items for which we do not have relevance judgments, i. e., they are assumed not to be relevant, but still potentially cover interests and tastes of any user, and

- when considering item features as aspects, we can consider as the set of aspects $\mathcal{A}$ either all the features or only those of the items in the user profile.

Agrawal et al. (2009) defined the family of so-called Intent-Aware metrics (see Section 2.4.1.2), in which state-of-the-art relevance-oriented metrics are adapted to measure the diversity of a search result list by aggregating the relative, marginalized relevance with respect to a set of *intents* or aspects of the query. The authors used ODP, a taxonomy of documents, to represent such query aspects. By replacing this taxonomy with our aspect space $\mathcal{A}$, we can naturally adapt such family of metrics to our recommendation scenario. In particular, the adaptation of the intent-aware version of the Expected Reciprocal Rank (ERR-IA) for recommendation tasks can be expressed as:

$$\mathrm{ERR} - \mathrm{IA}(R) = \sum_{a \in \mathcal{A}} p(a|u) \sum_{i \in R} \frac{1}{k_i} p(rel|i, u, a) \prod_{j \,:\, k_j < k_i} (1 - p(rel|j, u, a)) \tag{5.3}$$

where k_i is the position of item i in the recommendation list R and $p(rel|i, u, a)$ is the probability that the user u finds the recommended item i relevant when interested in aspect a. In the proposed feature-based aspect space, we consider the following, straightforward estimation of the probability $p(rel|i, u, a)$ by considering that an item is relevant with respect to a user u and a feature f as long as i contains f and the user finds the document relevant:

$$p(rel|i, u, f) = \mathbb{1}_{f \in \mathcal{F}_i}\, p(rel|i, u)$$

The relative importance of each user aspect is controlled in ERR-IA by the probability $p(a|u)$. In TREC, the subtopics of a query are often considered to be equally important in the absence of additional information – such as the intent mining method of Chapelle et al. (2011). As stated already in Section 5.2.2, this problem does not affect our feature-based aspects, where the user profile provides a reliable way of estimating the importance of each aspect.

In a similar direction, Clarke et al. (2008) (see Section 2.4.1.3) built on the notion of *information nuggets* (equivalent to interpretations and facets) to propose a framework to assess the *novelty* – in the sense of anti-redundancy – and *diversity* – in the sense of aspect coverage – of search results. Replacing such nuggets with user aspects, their proposed metric α-nDCG, which is a diversity-oriented re-formulation of the normalized Discounted Cumulative Gain of Järvelin and Kekäläinen (2000), results in the following metric for recommendations:

$$\alpha\text{-nDCG}(R) = \frac{1}{\alpha\text{-iDCG}} \sum_{i \in R} \left[\frac{1}{\log_2(k_i + 1)} \sum_{a \in \mathcal{A}} \mathrm{rel}(i \,|\, u, a) \prod_{j \,:\, k_j < k_i} (1 - \alpha \, \mathrm{rel}(j \,|\, u, a)) \right] \tag{5.4}$$

where $\mathrm{rel}(i \,|\, u, a)$ is a binary relevance judgment for item i given user u and aspect a. As in the case of ERR-IA, an immediate estimation for the case of feature-based aspects can be made by particularizing the relevance of an item i by its feature coverage and ad-hoc user relevance for item i:

$$\mathrm{rel}(i \,|\, u, f) = \mathbb{1}_{f \in \mathcal{F}_i} \, \mathrm{rel}(i \,|\, u)$$

Overall, we show here that our notion of user aspects allows us to adapt Information Retrieval diversity metrics to the recommendation setting in a principled manner. In particular, following the same principles other metrics for search result diversification are equally adaptable to recommendation, as it would be the case, for example, of the CPR metric (Dang and Croft, 2012).

5.3.2 *Adaptation of Diversification Methods*

As seen in Section 2.4.2, many search result diversification techniques (Carbonell and Goldstein, 1998; Agrawal et al., 2009; Santos et al., 2012; Dang and Croft, 2012) rely on a greedy re-ranking scheme in which documents from an initially retrieved result list are re-ranked by mean of an iterative procedure that selects, at each step, the document d that maximizes some objective function $f_{obj}(d \,|\, S)$ that measures the contribution to the diversity of the document when added to the set S of previously re-ranked documents.

In the context of Intra-List Diversity in recommendations, we present in Section 4.8 an equivalent greedy re-ranking strategy in which we iteratively select items from an initial set of recommended items by picking, at every step, the item that maximizes a linear combination between the original relevance scores – the one that defines the original ranking – and a novelty value with respect to the previously set S of already selected items:

$$s_{nov}(i \,|\, S) = (1 - \lambda) \, s(u, i) + \lambda \, nov(i \,|\, S) \tag{5.5}$$

In this section, we provide an adaptation of search result diversification methods by means of equating the objective function defined in search re-ranking and the novelty component in recommendation re-ranking, that is, $nov(i \,|\, S) = f_{obj}(i \,|\, S)$.

We start by considering the explicit Query Aspect Diversification framework (xQuAD) of Santos et al. (2010a). This framework presents a greedy re-ranking approach that can be cast to our recommendation setting by considering the probability for a user u selecting an item i but not the set S of already selected items:

$$\text{nov}_{\text{xQuAD}}(i\,|\,S) = p(i, \neg S\,|\,u) \tag{5.6}$$

In their approach, they marginalize this probability with query reformulations provided by a search engine as proxies for query aspects. In our setting, we make an adaptation by directly considering the user aspect space and substituting query reformulations with it:

$$\begin{aligned}
\text{nov}_{\text{xQuAD}}(i\,|\,S) &= p(i, \neg S\,|\,u) \\
&= \sum_{a \in \mathcal{A}} p(a\,|\,u)\; p(i, \neg S\,|\,u, a) \\
&= \sum_{a \in \mathcal{A}} p(a\,|\,u)\; p(i\,|\,u, a)\; p(\neg S\,|\,u, a) \\
&= \sum_{a \in \mathcal{A}} p(a\,|\,u)\; p(i\,|\,u, a) \prod_{j \in S} (1 - p(j\,|\,u, a))
\end{aligned}$$

where $p(i\,|\,u, a)$ is the probability of choosing item i given an aspect a of user u. Similarly as in previous estimations of components, we consider this probability to be proportional to the conjunction of feature coverage and general user selection probability for feature-based aspects:

$$p(i\,|\,u, f) = \frac{\mathbb{1}_{f \in \mathcal{F}_i}\; s(u, i)}{\sum_{j \in R} \mathbb{1}_{f \in \mathcal{F}_j}\; s(u, j)} \tag{5.7}$$

Agrawal et al. (2009) defined an alternative greedy diversification algorithm, called Intent-Aware Selection (IA-Select), to maximize the probability that the average user finds at least one relevant result within the retrieved results. Interestingly, by replacing their query intents – represented by a taxonomy of documents – with user aspects, the resulting formulation is actually equivalent to our adaptation of xQuAD when $\lambda = 1.0$.

The adaptation of the xQuAD diversification framework provides a practical and effective method of optimizing the diversity metrics adapted in Section 5.3.1. This adaptation is specially convenient when we considered the aspects defined by item features, in which case we work with the same aspect space in both diversification and evaluation phases, as opposed to most settings for search result diversification. However, we observe the following issues:

- The generative formulation of IA-Select and xQuAD imposes, in a sense, a model where the user selects a single document of the retrieved list. We wonder if a model based on maximizing the perceived relevance – rather than single document selection – would provide better diversification effects.

- Recommendations based on all user's preferences may not be fit to generate good recommendations for the varied interests or tastes of the user. For example, user preferences in classical music can be more useful than rock music favorites to recommend a classical music piece.

In the following sections, we address these points and propose two new diversification methods for recommendation settings.

5.4 EXPLICIT RELEVANCE MODELS

In this section, we propose an alternative formulation for the diversification of recommendations based on an explicit relevance model.

As Equation 5.6 shows, xQuAD – as well as other methods such as IA-Select – relies on the probability $p(i|u,a)$ of the user u interested in aspect a selecting the item i from the recommendation. This approach, in a sense, reflects a model of selection where the user selects a single item from the recommendation (Welch et al., 2011). In our alternative, we propose to model our diversification strategy in terms of the probability of relevance $p(\mathrm{rel}|i,u,a)$. Thus, we eliminate the conceptual restriction of selecting a single item in the recommendation.

We shall see that this approach has advantages of its own. it shows a competitive or better performance than its generative-based counterparts and, additionally, it allows further extensions and elaborations with models involving an explicit representation of relevance. As a particular case, we show that the framework provides a sound basis for tuning redundancy penalization in a principled way, as a smooth consistent extension of the diversity model.

5.4.1 *Relevance-Based xQuAD*

Initially, we reconsider the initial, generative-based formulation of the novelty component of the adaptation of the xQuAD framework for recommendation. Indeed, rather than expressing the novelty component in terms of the probability of selecting an item as in Equation 5.6, we initially define the novelty component on an explicit relevance model as $p(\mathrm{rel}_i, \neg\mathrm{rel}_S|u)$, where rel_i means i is relevant – that is, $p(\mathrm{rel}_i|u) = p(\mathrm{rel}|i,u)$ – and $\neg\mathrm{rel}_S$ means no document in S is relevant. Taking this starting point, by similar steps to the original xQuAD, we derive our relevance-based xQuAD (RxQuAD) as follows:

$$\begin{aligned}\mathrm{nov}_{\mathrm{RxQuAD}}(i\,|\,S) &= p(\mathrm{rel}_i, \neg\mathrm{rel}_S\,|\,u) \\ &= \sum_{a\in\mathcal{A}} p(a\,|\,u)\; p(\mathrm{rel}_i, \neg\mathrm{rel}_S\,|\,u, a) \\ &= \sum_{a\in\mathcal{A}} p(a\,|\,u)\; p(\mathrm{rel}\,|\,i, u, a)\; p(\neg\mathrm{rel}\,|\,S, u, a) \\ &= \sum_{a\in\mathcal{A}} p(a\,|\,u)\; p(\mathrm{rel}\,|\,i, u, a) \prod_{j\in S} (1 - p(\mathrm{rel}\,|\,j, u, a))\end{aligned} \tag{5.8}$$

where we have assumed rel_i and rel_S are conditionally independent given user u and aspect a.

To estimate the probability of relevance $p(\mathrm{rel}\,|\,i, u, a)$ in the case of feature-based aspects we consider, as in the previous adaptations, the combination of feature coverage and ad-hoc relevance. In particular, we adapt the idea of estimating relevance of (Chapelle et al., 2009) by mapping the predicted relevance – as given by the scores $s(u, i)$ – with an exponential function:

$$p(\mathrm{rel}\,|\,i, u, f) = \frac{2^{\mathbf{1}_{f\in\mathcal{F}_i}\, s(u,i)/s^*(u,f)} - 1}{2}$$

where $s^*(u, f)$ is the highest score given by the recommender to an item covering the feature f. In our previous publication (Vargas et al., 2012a), we took an alternative estimation based on the expected relevance at each rank position – obtained by performing a splitting on the training data –, which offered comparable results to those of the option presented here.

In the Section 5.6, we compare experimentally the performance of our RxQuAD method with the direct adaptation of xQuAD.

5.4.2 *Relevance-Based Redundancy Management*

Our relevance-based framework provides the basis for the introduction and derivation of further extensions on a formal probabilistic basis. We show this by extending our framework with an explicit model of the tolerance to redundancy: different tasks, or different users, introduce different conditions on how redundancy should be handled and penalized. We show next how this can be accounted for by a smooth generalization of our framework.

Let stop denote a binary random variable that is true when a user, in some recommendation list browsing context, stops exploring the recommendation. And let stop_S denote the fact that a user stops browsing after exploring some items in the subset S. We may refine the RxQuAD novelty component as $p(\mathrm{rel}_i, \neg\mathrm{stop}_S\,|\,u)$, where the marginal utility of the item i is defined in terms of the user stopping before reading i. This results into a nuanced reformulation of the objective function:

$$\begin{aligned} nov_{RxQuAD}(i \,|\, S) &= p(rel_i, \neg stop_S \,|\, u) \\ &= \sum_{a \in \mathcal{A}} p(a \,|\, u)\ p(rel \,|\, i, u, a) \prod_{j \in S} (1 - p(stop \,|\, j, u, a)) \end{aligned} \tag{5.9}$$

This form of the novelty function generalizes the original one by abstracting from the reasons why an item i – in the context of a particular recommendation list – would not add value to the effective utility of the recommendation list.

Now, we may marginalize the stopping probability with respect to relevance:

$$\begin{aligned} p(stop \,|\, j, u, a) &= p(stop \,|\, j, u, a, rel)\ p(rel \,|\, j, u, a) \\ &\quad + p(stop \,|\, j, u, a, \neg rel)\ (1 - p(rel \,|\, j, u, a)) \end{aligned}$$

where again different simplifications can be considered. First, within the objective function for greedy item selection, we should consider $p(stop \,|\, j, u, a, \neg rel) = 0$ for $j \in S$, as the utility of the next item (which the scoring function means to assess) would not be an issue if the user had stopped browsing already somewhere in S. Another reasonable simplification is to assume the user's decision to stop at a specific document only depends on finding relevance, i. e., $p(stop \,|\, j, u, a) = p(stop \,|\, rel)$, whereby the model reduces to:

$$p(stop \,|\, j, u, a) = p(rel \,|\, j, u, a)\ p(stop \,|\, rel)$$

This way the original diversification algorithm is generalized to a form where an additional parameter $p(stop \,|\, rel)$ represents the user tolerance to redundancy – or in some sense, how many items it takes for the user to be satisfied:

$$\begin{aligned} nov_{RxQuAD}(i \,|\, S) = \sum_{a \in \mathcal{A}} \Bigg[& p(a \,|\, u)\ p(rel \,|\, i, u, a) \\ & \prod_{j \in S} (1 - p(rel \,|\, j, u, a)\ p(stop \,|\, rel)) \Bigg] \end{aligned} \tag{5.10}$$

The introduction of this additional parameter allows to better match this characteristic of users and/or recommendation scenarios. It allows to control (raise or soften) the penalization that should be applied to items relevant to aspects already covered in the ranking. The basic xQuAD and RxQuAD formulations implicitly assume $p(stop \,|\, rel) = 1$, that is, the user stops as soon as he finds a relevant document (zero tolerance to redundancy), which reflects again an implicit assumption that users are willing to select a single document – which is often not the case.

An equivalent parameter might be inserted in the original xQuAD formulation to soften redundancy penalization, but it would lack the formal justification that the relevance-based approach enables. Furthermore, the xQuAD redundancy penalization is already rather mild compared to RxQuAD, since the discounting term of the novelty component is based on item probabilities $p(i|u,a)$, which tend to range on much lower values (since they should sum to 1 over all items covering an aspect) compared to a Bernoulli relevance distribution $p(\mathrm{rel}|i,u,a)$. The addition of a tolerance parameter to xQuAD would only make this worse – unless it ranged beyond $[0,1]$, which would bring the scheme even farther from a formal probabilistic basis.

On the other hand, tolerance to redundancy has also been explicitly modeled and introduced in the context of metric formalization upon user models (Carterette, 2011; Clarke et al., 2011a; Hu et al., 2011). Therefore the use of this parameter in our diversification algorithm has the potential of a better optimization for such metrics by bringing the diversification model closer to the principles and assumptions which are built into the metrics.

A part of our experiments, we show a confirmation of the soundness of the redundancy management of RxQuAD when considering different degrees of tolerance to redundancy in the evaluation metric α-nDCG.

5.5 USING USER SUB-PROFILES

The estimation for the probability $p(i|u,a)$ as done in Equation 5.7 allows to effectively diversify recommendations. This approach, however, relies on the assumption that the items covering a particular taste or interest of the user in a recommendation generated by using the whole profile of the user – that is, representing the diversity of interests of the user – adequately represents the preferences of this user for that particular interest. In this section, we propose an alternative beyond this assumption by embracing the diversity of user preferences in the recommendation process itself.

We consider and adapt the idea of Santos et al. (2010a) of using query reformulations as proxies for multiple interpretations or aspects of an ambiguous or underspecified query. Consider an ambiguous query like “java” and some reformulations like “java island”, “java coffee” or “java programming”, which specify or disambiguate to some extent the original query. Such reformulations can be obtained from commercial search engines, and can serve as a proxy of query intents. The results obtained for the reformulations are expected to better answer the specific intents underlying the original query, so that a combination of them – by means of the xQuAD scheme – may result in a diversified search result that copes with the ambiguity and underspecification of the original query “java”.

In our proposal, we "re-formulate" user profiles by considering subsets of it representing a single interest or taste, which we call **sub-profiles**. Our assumption is that the recommendations generated exclusively with the information in each sub-profile better reflect the specific preferences of the user for a given taste. Then, by adapting the greedy selection of the xQuAD and RxQuAD schemes, we can compose diversified recommendation lists that adequately capture the heterogeneous preferences of the users. Our approach is thus composed of three steps: extraction of sub-profiles, generation of recommendations for these sub-profiles, and combination of these recommendations into single, diverse recommendations representing the different tastes of the user.

The work of Zhang and Hurley (2009) explores a notion of user profile partitioning which can be related to our proposal, with significant differences nonetheless:

- In our approach, user profiles are partitioned using a previously available categorization of the item domain, whereas their approach requires more elaborate clustering algorithms to define their partitions based on the similarity between user ratings.

- In their work, once partition-specific recommendations have been generated, a selection of the most novel ones is combined into a final recommendation by uniformly allocating items from each partition-specific recommendations. We propose a combination of sub-profile recommendations by means of our non-trivial adaptation of xQuAD, which performs a rank-aware allocation of items by taking into account the relative importance of each sub-profile while maximizing the number of user tastes represented while avoiding redundancy in the recommendation.

Given a user $u \in \mathcal{U}$ with profile $\mathcal{I}_u$, we first define and generate every user sub-profile $\mathcal{I}_u^a \subset \mathcal{I}_u$ as the subset of the original profile that represents an aspect $a \in \mathcal{A}$ reflecting a particular interest or taste of the user. In the case of item feature-based aspects, i. e., $\mathcal{A} = \mathcal{F}$, the sub-profiles are straightforwardly defined by the features of the items:

$$\mathcal{I}_u^f = \{i \in \mathcal{I}_u \, : \, f \in \mathcal{F}_i\}$$

Conceptually, we may think of a sub-profile $\mathcal{I}_u^a$ as representing some abstract *sub-user* u^a which has a unique, clearly defined interest or taste.

The next step consists in generating recommendations for each sub-profile or, conceptually, for the *sub-users* these sub-profiles represent. Collaborative filtering algorithms generate recommendations for users by combining preferences of similar users. In our setting, the newly defined *sub-users* are used for the purpose of generating recommendations for other *sub-users* by means of collaborative filtering algorithms. Formally, our collaborative filtering setting for sub-profiles considers a community of users defined as $\mathcal{U}^{\mathcal{A}} = \bigcup_{a \in \mathcal{A}} \{u^a\}_{u \in \mathcal{U}}$ where the interaction between the users in $\mathcal{U}^{\mathcal{A}}$ and the items $\mathcal{I}$ is defined in an interaction matrix $\mathcal{R}^{\mathcal{A}}$ derived from the original matrix $\mathcal{R}$ and the sub-profiles of the *sub-users*.

The final step consists in combining the recommendations of the sub-profiles of a user to create a single and diverse recommendation list. This is done by adapting the greedy scheme of xQuAD (Equation 5.6) to consider sub-profiles. Specifically, we replace the probability $p(i|u, a)$ by $p(i|u^a)$, which represents the likelihood of the item i being selected by *sub-user* u^a, that is, the abstract user defined by the sub-profile $\mathcal{I}_u^a$ representing a drive (a sub-taste of the original user) for the taste or interest represented by aspect a. We estimate this probability $p(i|u^a)$ as proportional to the score $s(u^a, i)$ in the recommendation R_{u^a} assigned to the item i for the *sub-user* u^a:

$$p(i|u^a) = \frac{s(u_a, i)}{\sum_{j \in R_{u_a}} s(u_a, j)} \tag{5.11}$$

The resulting diversification method, which we call SxQuAD, is defined as follows:

$$\begin{aligned} nov_{SxQuAD}(i|S) &= p(i, \neg S|u) \\ &= \sum_{a \in \mathcal{A}} p(u_a|u)\, p(i|u_a) \prod_{j \in S} (1 - p(j|u_a)) \end{aligned} \tag{5.12}$$

where $p(u_a|u) = p(a|u)$.

Similarly, we can also adapt the RxQuAD of Section 5.4 (Equation 5.8) to use sub-profiles. We do this by replacing the probability $p(rel|i, u, a)$ by $p(rel|i, u^a)$. We call the resulting variant SRxQuAD:

$$\begin{aligned} nov_{SRxQuAD}(i|S) &= p(rel_i, \neg rel_S|u) \\ &= \sum_{a \in \mathcal{A}} p(u_a|u)\, p(rel|i, u_a) \prod_{j \in S} (1 - p(rel|j, u_a)) \end{aligned} \tag{5.13}$$

The estimation of the probability of relevance for the item i and sub-user u^a is given by the following formula:

$$p(rel|i, u^a) = \frac{2^{s(u^a,i)/s^*(u^a)} - 1}{2} \tag{5.14}$$

where $s^*(u^a)$ is the highest score in the recommendation R_{u^a} for the sub-profile targeting aspect a.

Both resulting methods from considering sub-profiles are being compared with the direct adaptation of xQuAD and the previously presented relevance-based RxQuAD in the next section.

5.6 EXPERIMENTS

Following again the general experimental design described in Chapter 3, we have conducted an experimental evaluation to show the consistency of our adaptation of the Intent-Aware framework to Recommender Systems and the enhancements provided by our novel proposals to the diversity of recommendations under this framework. Concretely, our experiments aim to answer the following questions:

- What is the performance of state of the art baselines when measured with our adapted metrics? Can they be improved by means of our adapted diversification methods?
- Are our newly-proposed diversification proposals (RxQuAD, SxQuAD and SRxQuAD) able to improve over the direct adaptations of the methods for search result diversification?
- What is the effect of the relevance-based redundancy management of Section 5.4.2?

As suggested throughout this chapter, we defined our aspect space using features of the items in the user profiles, specifically genres as detailed in Section 3.2. Based on this aspect space, we measured the performance of baseline recommendation algorithms and their diversifications with our adaptations of ERR-IA (Equation 5.3), α-nDCG (Equation 5.5) and S-recall (Equation 5.2), this last one without taking into account the relevance of the recommended items, just the aspects – in this case genres – covered by all the recommended items. We also used simple precision to compare with ERR-IA and α-nDCG ($\alpha = 0.5$), which have been claimed to strongly correlate with ad-hoc relevance when deployed in search tasks (Golbus et al., 2012). All metrics were evaluated at cut-off 20.

5.6.1 *Evaluation of Baseline Recommendation Algorithms*

In Table 5.1 we can see the results of measuring the baseline recommendation algorithms detailed in Section 3.3 in our three datasets with our adapted Intent-Aware metrics using genres as aspects. The results indicate that both ERR-IA and α-nDCG are clearly oriented towards measuring the "pure" relevance of the results, although they are also able to reflect the inherent diversity of the recommendation algorithms. We see this by analyzing separately relevance and diversity by means of the precision and relevance-unaware S-recall, respectively.

We first analyze the results in the MovieLens1M dataset. In terms of S-recall, we can see that the most-popular recommendation offers a very high number of genres. This result coincides with the findings in Table 4.2, in which we observed

		P	**ERR-IA**	**α-nDCG**	**S-recall**
ML1M	**Rnd**	0.0057	0.0044	0.0085	0.6691
	Pop	0.1215	0.0962	0.1942	0.7206
	iMF	0.2335	0.2212	0.3756	0.6621
	pLSA	0.2111	0.1918	0.3315	0.6542
	UB	0.2055	0.2049	0.3523	0.6872
	IB	0.1874	0.1832	0.3121	0.6445
Netlfix	**Rnd**	0.0022	0.0014	0.0031	0.6025
	Pop	0.0909	0.0670	0.1406	0.4091
	iMF	0.1778	0.1412	0.2656	0.4861
	pLSA	0.1842	0.1429	0.2640	0.4756
	UB	0.1923	0.1562	0.2952	0.4888
	IB	0.1582	0.1250	0.2418	0.4739
MSD	**Rnd**	0.0001	0.0000	0.0001	0.7507
	Pop	0.0185	0.0098	0.0374	0.7701
	UB	0.1018	0.1034	0.2221	0.6107
	IB	0.1078	0.0959	0.2084	0.5285

Table 5.1: Results of the adapted Intent-Aware metrics for the baseline recommendation algorithms in MovieLens1M, Netflix and Million Song Dataset.

how the most-popular recommendation achieved a high Expected Intra-List Distance when measured by genres. Again, we justify this effect by the variety of tastes among users. Surprisingly, the second best performing baseline in terms of S-recall is the user-based nearest neighbors algorithm, which performs better than random recommendations and the rest of personalized algorithms. As pointed out before, ERR-IA and α-nDCG are biased towards measuring the relevance of the recommendations. In particular, their results do not differ, in relative terms, to those of the simple diversity-insensitive precision, with the exception of the user-based nearest neighbors method. Although pLSA baseline offers a slightly better accuracy (in terms of precision) than UB, it clearly has a lower recall of genres that UB (as S-recall points out), which results in higher values of ERR-IA and α-nDCG for the UB baseline.

The outcomes with the Netflix show a similar trend to that of MovieLens1M. In this case, the best performing algorithm in terms of S-recall is the random one, while the worst is popularity. The low performance of the most-popular recommendation contrasts with our expectations given its relatively good performance in terms of $\mathrm{EILD}_{\mathcal{F}}$ in Table 4.2. This result indicates however that both ways of measuring Intra-List Diversity do not necessarily agree completely. The personalized recommendation algorithms obtain comparable results for S-recall, showing a similar relevance-unaware diversity, which probably explains why ERR-IA and α-nDCG do not differ from the outcomes from precision.

Finally, in the Million Song Dataset both the random and most-popular recommendations get the highest results in terms of genre recall. The results for S-recall of both personalized recommendations are quite different and explain the outcomes of both ERR-IA and α-nDCG: although having similar performance in terms of precision, the higher recall of genres of the user-based approach favors this algorithm when compared to the item-based approach in ERR-IA and α-nDCG.

5.6.2 *Evaluation of Re-Ranking Strategies*

Concerning the second question that drives our experiments, we applied four different re-ranking strategies to the baselines of the MovieLens1M dataset: the direct adaptation of the xQuAD framework to recommendations, its explicit relevance-based version RxQuAD, and the sub-profile-based variants of these, namely the generative-based SxQuAD and the relevance-based SRxQuAD, respectively. All these greedy re-ranking strategies are controlled by a λ parameter, for which we performed a grid search from $\lambda = 0.0$ (no diversification) to $\lambda = 1.0$ (full diversification) with steps of 0.1.

The results of the four diversification strategies applied on the personalized baseline recommendation algorithms in MovieLens1M are shown in Figure 5.1, Figure 5.2 and Table 5.2. The figures show the performance in terms of the four considered metrics – precision and the Intent-Aware adapted metrics – for every explored value of λ for each recommendation algorithm. Table 5.2 shows, for the sake of summarization, the results for the value of λ that optimizes ERR-IA for each baseline and re-ranking diversification method. In global terms, these results illustrate the properties of our adapted Intent-Aware framework, particularly the consistency and the effectiveness of our proposed methods.

As a general trend, we observe that all the re-ranking diversification strategies imply a certain trade-off between accuracy and diversity. That is particularly manifest in the results of precision and S-recall: the higher the λ, the lower the precision – with the exception of the sub-profile methods – and the higher the S-recall – with the exception, again, of some results of the sub-profile methods differing from the global trend. In the case of ERR-IA and α-nDCG, this trade-off is reflected in the fact that intermediate values of λ achieve the best results in terms of these metrics for all considered approaches and, depending on the choice of baseline and diversification strategy, a full diversification ($\lambda = 1.0$) can result in improvements with respect to the original baseline even when the precision is negatively affected.

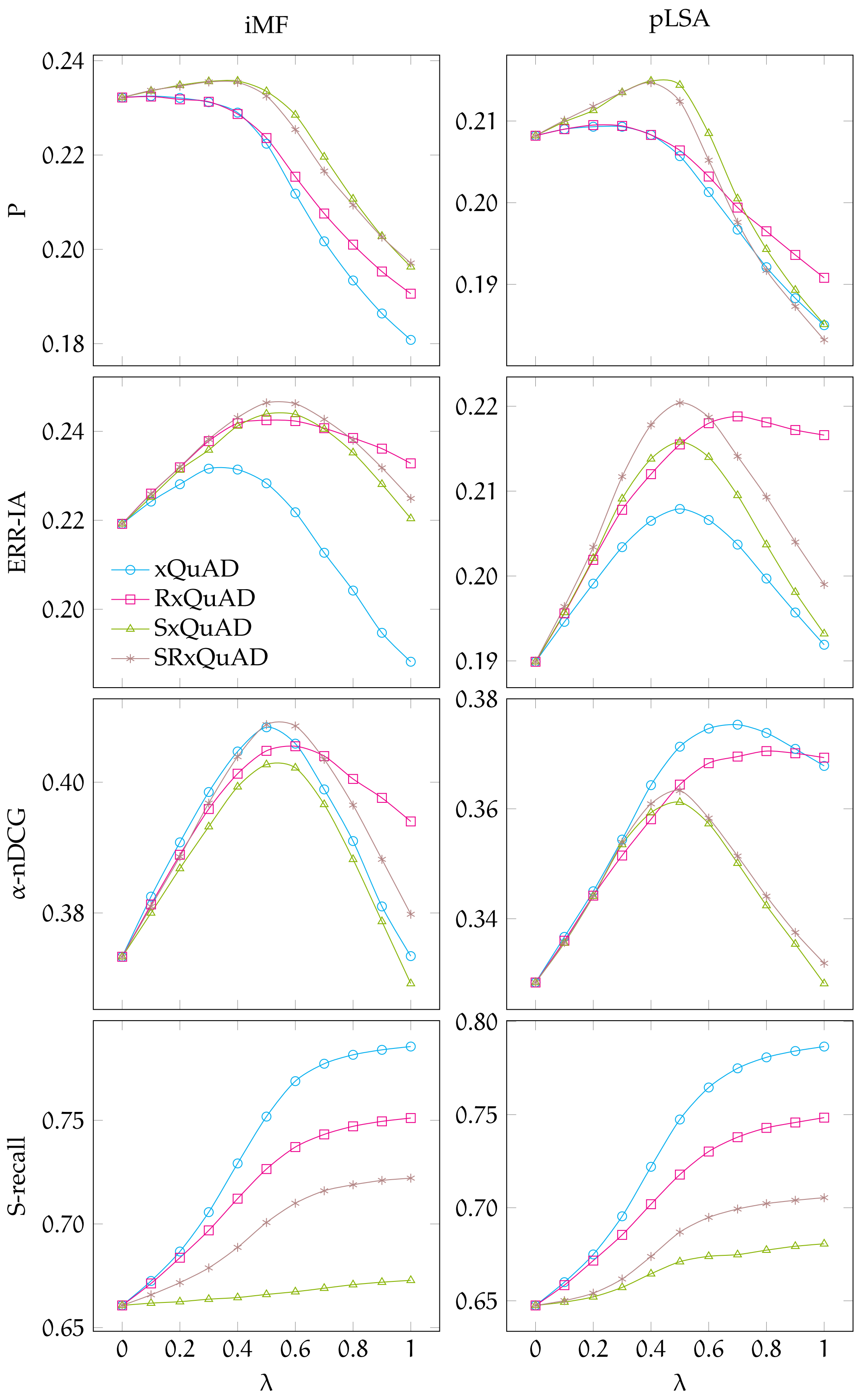

Figure 5.1: Intent-Aware re-rankings for the latent factors recommendation algorithms in MovieLens1M.

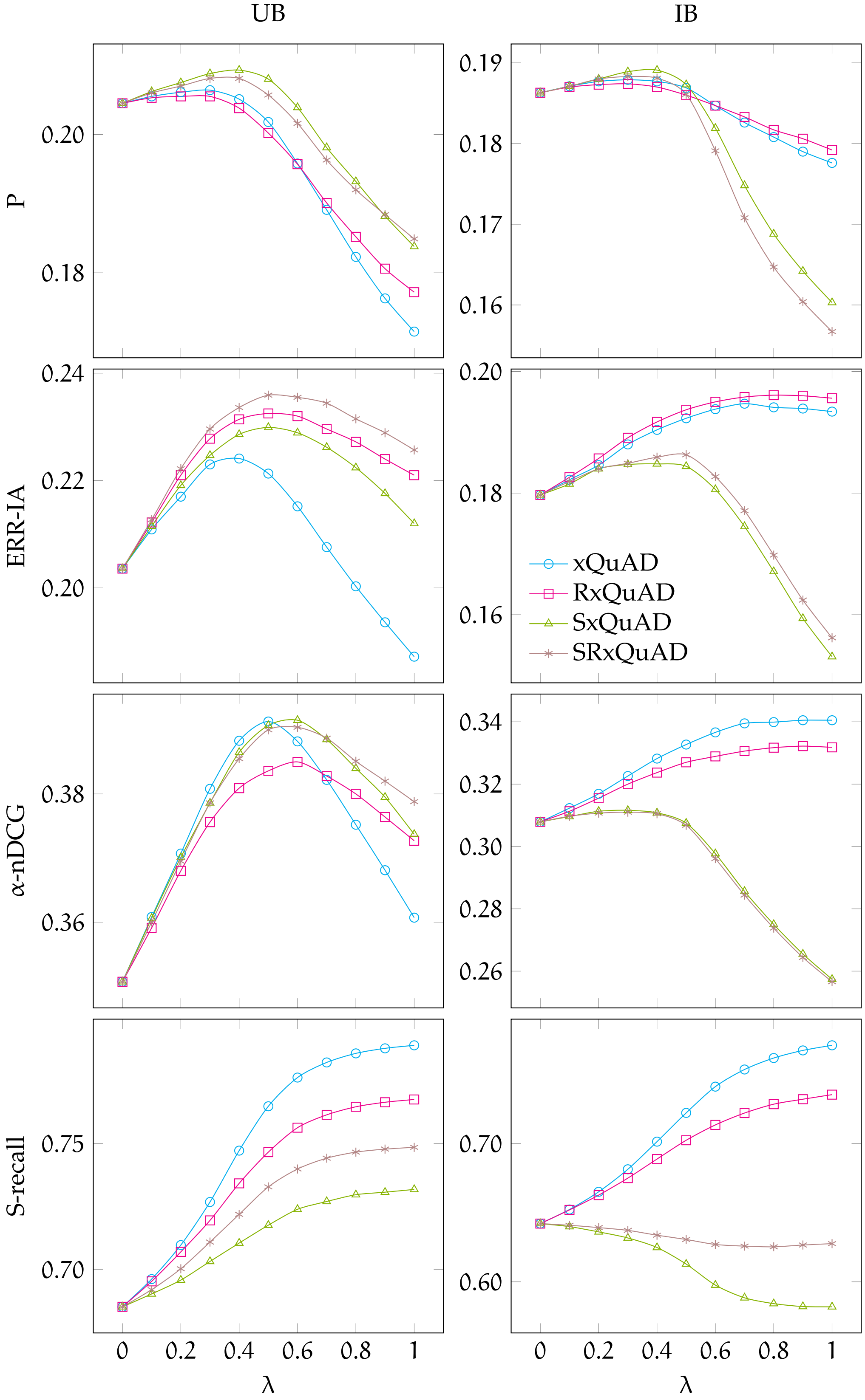

Figure 5.2: Intent-Aware re-rankings for the nearest neighbors recommendation algorithms in MovieLens1M.

	P	ERR-IA	α-nDCG	S-recall
iMF	0.2335	0.2212	0.3756	0.6621
+**xQuAD** (0.4)	0.2301	0.2334	0.4076	0.7299
+**RxQuAD** (0.6)	0.2162	0.2453	0.4081	0.7382
+**SxQuAD** (0.6)	0.2297	0.2460	0.4036	0.6684
+**SRxQuAD** (0.5)	0.2335	0.2485	0.4099	0.7012
pLSA	0.2111	0.1918	0.3315	0.6542
+**xQuAD** (0.5)	0.2077	0.2099	0.3723	0.7501
+**RxQuAD** (0.7)	0.2013	0.2201	0.3718	0.7394
+**SxQuAD** (0.5)	0.2161	0.2204	0.3663	0.6746
+**SRxQuAD** (0.5)	0.2150	0.2238	0.3676	0.6924
UB	0.2056	0.2049	0.3523	0.6872
+**xQuAD** (0.4)	0.2067	0.2260	0.3903	0.7488
+**RxQuAD** (0.5)	0.2015	0.2341	0.3859	0.7480
+**SxQuAD** (0.5)	0.2091	0.2314	0.3926	0.7190
+**SRxQuAD** (0.5)	0.2069	0.2374	0.3914	0.7342
IB	0.1874	0.1832	0.3121	0.6445
+**xQuAD** (0.7)	0.1844	0.1997	0.3439	0.7555
+**RxQuAD** (0.8)	0.1831	0.2019	0.3366	0.7302
+**SxQuAD** (0.4)	0.1900	0.1881	0.3136	0.6269
+**SRxQuAD** (0.4)	0.1894	0.1895	0.3136	0.6359

Table 5.2: Detail from Figures 5.1 and 5.2 for the value of λ that optimizes ERR-IA for each baseline and re-ranking diversification strategy in MovieLens1M.

Regarding the outcomes of the different diversification strategies, we see that our proposed diversification strategies – RxQuAD, SxQuAD, and SRxQuAD – are able to improve over our simple, initial adaptation of the xQuAD framework for recommendations in terms of ERR-IA and α-nDCG, but not when measured by S-recall. The results depend though on the choice of the diversified recommendation baseline. In the case of the implicit matrix factorization, we see that our three proposed methods clearly improve over xQuAD in ERR-IA. The comparison results harder when measured by α-nDCG, in which case only the RxQuAD stands out by offering the best result for fully diversified baselines, i.e., when $\lambda = 1.0$. When we analyze the probabilistic Latent Semantic Analysis baseline, we see an unexpected positive increase in precision by the sub-profile methods. Again, in terms of ERR-IA, our proposals consistently improve over xQuAD. The good

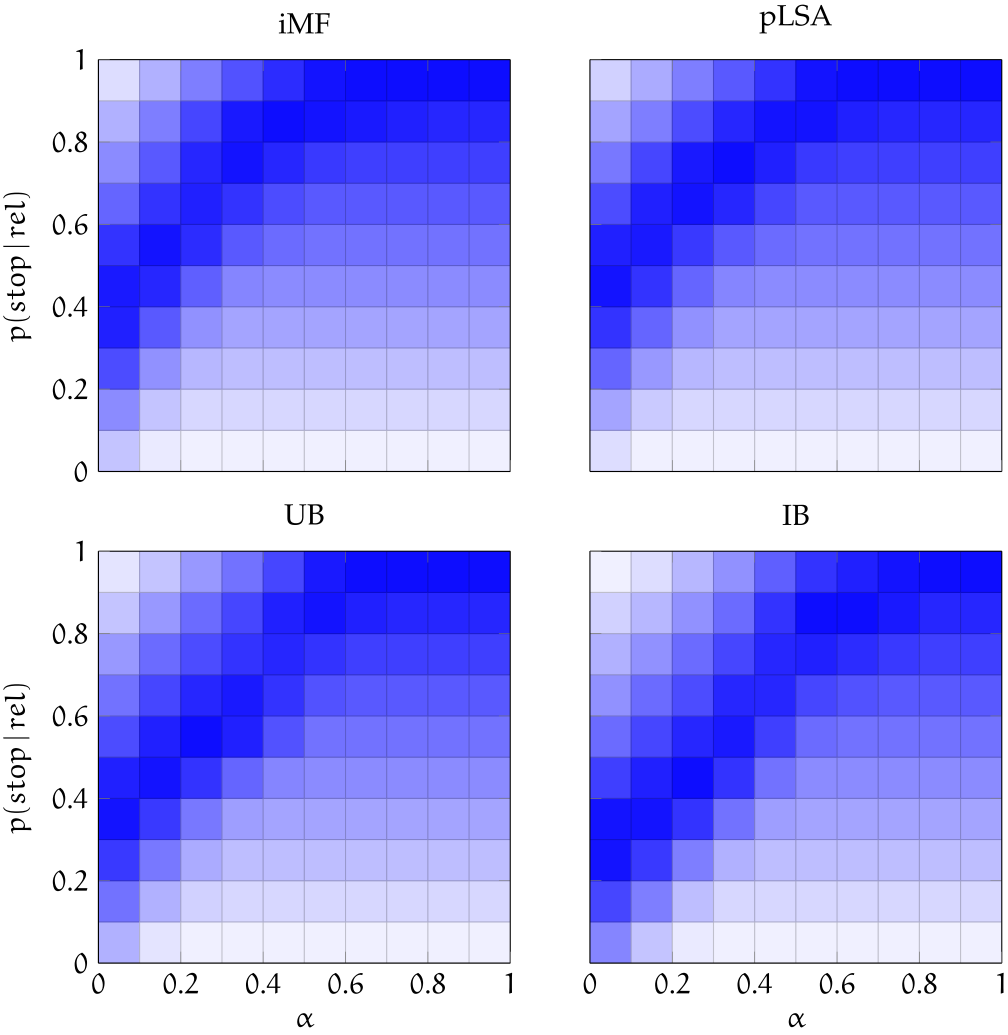

Figure 5.3: Parameterized tolerance to redundancy in the RxQuAD diversification framework by $p(\mathrm{stop} \mid \mathrm{rel})$ evaluated with α-nDCG for the personalized recommendation baselines in MovieLens1M. The values are displayed as a heat map where colder colors (rank-normalized per column) represent higher α-nDCG values.

trade-off of RxQuAD particularly stands out as it approaches a full diversification ($\lambda = 1.0$). Again, α-nDCG establishes a tough evaluation criterion that clearly benefits xQuAD. The results of the user-based nearest neighbors agree with the previous algorithms: our proposals perform better than xQuAD in ERR-IA, but according to α-nDCG such improvements are not so clear. Finally, the item-based nearest neighbors is the worst baseline for our approaches, specially for the sub-profile-based methods. As we can see, RxQuAD offers in this case, at best, a similar performance to that of the xQuAD method, while the sub-profile methods cannot even improve the S-recall of the original recommendations and, therefore, achieve a very poor performance in terms of the rest of the metrics.

In general terms, we see that our approaches are competitive when compared with a direct adaptation of the xQuAD method to recommendation. However, when compared against each other, the evidence of the previous results favors the simpler, more robust RxQuAD which, with a simpler formulation, is comparable to the sub-profile methods there where they perform the best, and performs relatively well where these are not so advantageous.

5.6.3 *Evaluation of the Redundancy Management of RxQuAD*

In order to illustrate the effect of the adjustable redundancy of Section 5.4.2, we display as a heat maps in Figure 5.3 the performance values of the generalized RxQuAD with different values of $p(\mathrm{stop}\,|\,\mathrm{rel})$, measured with α-nDCG with different values of α (also reflecting different degrees of redundancy tolerance, see Section 2.4.1.3) in the MovieLens1M dataset for all four personalized recommendation baselines. For each combination of $p(\mathrm{stop}\,|\,\mathrm{rel})$ and α, we select the λ parameter in RxQuAD that achieves the best results with respect to the evaluated metric. It can be observed that the redundancy penalization effect of $p(\mathrm{stop}\,|\,\mathrm{rel})$ is consistent with the equivalent parameter in the metric, i. e., the values evolve on a diagonal pattern: higher $p(\mathrm{stop}\,|\,\mathrm{rel})$ values in the algorithm perform better for higher α in the metric, and vice versa.

5.7 conclusions

In this chapter we have presented a principled adaptation of the metrics and methods of the Intent-Aware framework for Information Retrieval diversity to Recommender Systems. Our adaptation is based on the relationship between the motivation of search result diversification, namely query ambiguity and underspecification, and the inherent ambiguity of the user needs in the recommendation problem. We define the concept of user aspect as the analog of query intents, interpretations, facets or subtopics in the search problem. Based on these user aspects, we adapt several well-known metrics and diversification methods in Information Retrieval diversity to Recommender Systems. Moreover, we propose two new re-ranking diversification strategies on top of our adapted framework that obtain competitive or better results than directly adapted methods from the state of the art. An empirical evaluation supports the consistency of our adaptation and shows the effectiveness of different re-ranking diversification methods.

6

COVERAGE, REDUNDANCY AND SIZE-AWARENESS IN GENRE DIVERSITY FOR RECOMMENDER SYSTEMS

6.1 INTRODUCTION

The study of Intra-List Diversity in Recommender Systems has occupied an important part of this thesis. We review in Chapter 2 several existing approaches for the study of this quality dimension of recommendations that addresses the diversity of user interests and tastes and the user's need for more varied recommendations. In Chapter 4 we propose a distance-based measurement of this perspective that generalizes the Intra-List Distance of Smyth and McClave (2001) by considering rank and relevance of the evaluated recommendations. An adaptation of the Intent-Aware framework of search result diversification to Recommender Systems is proposed in Chapter 5 to address the diversity problem from a perspective that considers the ambiguity given by the heterogeneity of tastes of user profiles. Several of these prior approaches rely on the features of the items in the recommendation domain to measure the diversity of recommendation lists. We assume that features such as categories, genres, tags, etc. are a reasonable and effective source for estimating, for example, the distance (as the complement of similarity) between items and to determine the diverse interests or tastes of the user in a recommendation domain. In particular, in the experiments in movie and music domains we rely on genres as commonly accepted and reliable features available in such recommendations domains.

In this chapter, we delve into the specific case of defining diversity in recommendation by means of the genres of the items. As opposed to the prior approaches, where genres were conveniently but circumstantially used to determine the distances between items and to represent the user interests and tastes – as a replacement of subtopics –, we consider now the problem of providing diverse recommendations expressly by means of the genres available in recommendation domains such as movies, music or books. For this purpose, we analyze the properties of genres and their utility in providing diverse recommendations. We postulate three important properties that genre-based diverse recommendations should fulfill:

- **genre coverage**, that is, each genre should be represented in a recommendation list according to both the interest of the user and its specificity;

- **redundancy**: while it is important that all genres are represented it is equally important not to over-represent a particular genre – this is particularly important in domains where items can have more than one genre; and
- recommendation **list size-awareness**, which focuses on the common screen space limitation to offer recommendations, and how it influences genre coverage and redundancy.

Our analysis of state of the art diversification methods and metrics shows that they do not properly or fully address these three properties when they consider genres as a source of diversity. We propose in this chapter a new **Binomial** diversity framework that takes into account all the aforementioned properties. The framework consists of a metric to assess the diversity of recommendations and a greedy re-ranking strategy to optimize the diversity of recommendations. We report experiments in the context of our experimental design of Chapter 3 showing the properties of our framework, and comparing it to state of the art methods.

The rest of the chapter is structured as follows. In Section 6.2 we briefly remind the current state-of-the-art techniques for modeling recommendation diversity that our Binomial framework compares to. Section 6.3 presents the characteristics of genres and provides arguments for their use as a source of diversity in recommendations. We elaborate on the properties that genre-based diversity approaches should fulfill in Section 6.4. Section 6.5 proposes a framework for both evaluating and enhancing the genre-based diversity of recommendations. In Section 6.6 we conduct an experimental evaluation to show the validity of our approach compared with prior well-known approaches. Finally, Section 6.7 offers the discussion and conclusions.

The contents of this chapter have been presented in following published work:

- Vargas, S. and Castells, P. (2014a). Improving sales diversity by recommending users to items. In *Proceedings of the 8th ACM Conference on Recommender Systems*, RecSys '14, pages 145–152, New York, NY, USA. ACM
- Vargas, S., Castells, P., and Vallet, D. (2012b). On the suitability of intent spaces for IR diversification. In *Proceedings of the International Workshop on Diversity in Document Retrieval at the 5th ACM International Conference on Web Search and Data Mining*, DDR'12, Seattle, Washington, USA

6.2 COMPARED APPROACHES

Different frameworks for measuring and enhancing Intra-List Diversity have been proposed in the Recommender Systems and Information Retrieval literature. We briefly recall here the most closely related and relevant research to the scope of this chapter.

One of the earliest and best-known proposals for diversity in Recommender Systems is the Intra-List Distance of Smyth and McClave (2001), which we generalize in Chapter 4 with our Expected Intra-List Distance (EILD) metric and its corresponding greedy re-ranking strategy. We will refer to this approach as the **Pair-Wise framework** throughout this chapter. This framework takes a distance measurement between pairs of items as a basis to determine the diversity of recommendation lists. Given a set of genres $\mathcal{G}$, we propose to measure the distance as the complement of the cosine similarity of the genres covered by each item:

$$\mathrm{sim}(i,j) = \frac{|\mathcal{G}_i \cap \mathcal{G}_j|}{\sqrt{|\mathcal{G}_i|\,|\mathcal{G}_j|}}$$

Another major line of work in measuring and enhancing diversity in search result diversification, the **Intent-Aware framework**, was adapted in Chapter 5 to the recommendation setting. Given a set of genres, they serve as a basis to extract the user aspects – analogous to query subtopics in search result diversification – that represent the different interests and tastes of the users. For example, the ERR-IA metric with genres playing the role of user aspects is instantiated as follows:

$$\mathrm{ERR-IA}(R) = \sum_{g \in \mathcal{G}} p(g|u) \sum_{i \in R} \frac{1}{k_i}\, p(rel|i,u,g) \prod_{j\,:\,k_j<k_i} (1 - p(rel|j,u,g))$$

Similarly, the rest of the metrics (S-recall, α-nDCG) and diversification methods (IA-Select, xQuAD) can be used with genres, as seen in Section 4.9.

A third approach is the more recent **Proportionality framework** by Dang and Croft (2012) for search result diversification reviewed in Section 2.4.1.4. This framework emphasizes the need for covering each subtopic of the search query by offering a number of relevant documents proportional to the interest of the subtopic they cover. For the purpose of this chapter, we propose an adaptation of this framework to the recommendation problem in a similar manner to that of the Intent-Aware framework in Chapter 4. The basis for measuring this framework is the Disproportionality (DP) or square loss of the under-represented genres in a recommendation list:

$$\mathrm{DP}(R) = \sum_{s} \mathbf{1}_{v_g \geqslant k_g} (v_g - k_g)^2$$

where v_g and k_g are the expected and actual numbers of items that cover the genre g in the recommendation R respectively. On top of DP, Dang and Croft propose a Cumulative Proportionality metric (CPR) that is the basis of their study. Likewise, we also adapt the Proportionality Method (PM, see Section 2.4.2.2) re-ranking strategy, which is inspired on a seat assignment system for legislative elections in some countries.

Genre	Count	Genre	Count	Genre	Count	Genre	Count
Action	517	Crime	329	Horror	368	Thriller	600
Adventure	387	Documentary	105	Musical	137	War	141
Animation	107	Drama	1,711	Mystery	154	Western	71
Children	258	Fantasy	181	Romance	583		
Comedy	1,267	Film-Noir	49	Sci-Fi	283		

Table 6.1: Genre distribution in Movielens1M.

6.3 CHARACTERIZING GENRES

As defined in the Merriam-Webster dictionary[1], a genre is "a category of artistic, musical, or literary composition characterized by a particular style, form, or content". We argue that genres can be used as the source for defining diversity as they:

- explicitly define a conventional style of an item that has a common interpretation among users,
- have the potential of representing the different tastes of individual users,
- are well accepted for media categorization and are already available in most online media catalogs for movies, literature, music, etc., and
- it is safe to assume that the user will perceive the diversity of the recommendation list if the genres are diversified among the recommended items. Other alternatives such as using item-to-item distance based on consumption patterns may have an effect on the inherent diversity of the recommendation, although this may not directly translate to a user perception of diversity.

Genres, nonetheless, present some particularities that need to be addressed to be used effectively. First, genres can have different levels of generality: for example, in the movie domain "Drama" represents a very broad and vaguely defined style with many diverse movies belonging to this genre. On the other hand, "Western" is a quite specific movie type which is usually devoted to telling stories in the American Wild West. This generality is also reflected in the number of items for each genre: more general genres will usually be present in higher number of items than more specific genres. We observe that the generality of each genre is also related to the perception of redundancy in a recommendation list. For example, three random westerns in a short recommendation list of five items feels more redundant than three random dramas. We will exploit this observation when defining our genre-based diversity model. Second, genres do not usually define disjoint

[1] http://www.merriam-webster.com

Genre	Count	Genre	Count	Genre	Count	Genre	Count
Action	1,464	Documentary	779	Horror	900	Sci-Fi	819
Adult	54	Drama	4,408	Music	568	Short	237
Adventure	996	Family	772	Musical	418	Sport	284
Animation	381	Fantasy	651	Mystery	709	Talk-Show	2
Biography	384	Film-Noir	70	News	1	Thriller	1,989
Comedy	3,025	Game-Show	2	Reality-TV	15	War	422
Crime	1,319	History	317	Romance	1,887	Western	285

Table 6.2: Genre distribution in Netflix.

or isolated categories in their domains, and it is generally difficult to establish a precise hierarchy among them. For example, “The Lord of the Rings” by J. R. R. Tolkien can be classified as Adventure, Fiction, High fantasy and British literature all at once. Moreover, careless use of sub-genres can lead to lower perceived diversity. For example, heavy metal and white metal – two closely related sub-genres – share the same musical techniques, modes of dress and performance and could be perceived as similar by a listener.

These properties contrast with the characteristics of query aspects in search result diversification, specially as defined in the *subtopic retrieval* problem (Zhai et al., 2003) that has motivated the different diversity tasks in the TREC Web track (Clarke et al., 2009, 2010, 2011b, 2012). First, there is no notion of subtopic generality as they are defined uniquely for each query. Second, subtopic overlaps are expected to be much less frequent, specially in the case of ambiguous queries, in which the different interpretations are naturally covered by (mostly) disjoint sets of documents.

In order to study the properties of genres, we analyze the case of the datasets in our experimental design of Chapter 3 in movie and music recommendation. First, as previously stated, the generality of genres is manifested in the number of items covering each of them. We therefore illustrate this property by counting the number of items in our datasets covering each genre. Second, in order to show how genres overlap between each other without any hierarchical pattern, we show the overlap of the five most frequent genres in each domain in the form of Venn diagrams and the distribution of the number of genres each item covers. The resulting tables and figures properly confirm the previously enunciated particularities of genres and provide a justification for our requirements for measuring genre-based diversity.

Table 6.1 and Figure 6.1 show the statistics of genres for the MovieLens1M dataset. As the table shows, the number of movies for each genre varies greatly, from 1,711 movies in “Drama” to only 49 in “Film-Noir”. As we can observe, nar-

Genre	Count	Genre	Count	Genre	Count
Alternative Rock	28,246	Folk	14,081	Metal	22,799
Ambient	10,476	Funk	6,428	Pop Rock	5,862
Blues	7,743	Hard Rock	11,337	Punk	18,054
Classic Rock	13,840	Hardcore	8,768	Rap	13,641
Country	8,117	House	5,221	Reggae	5,354
Dance	15,803	Indie Pop	8,593	Soul	10,659
Electronica	13,823	Jazz	16,063	Trance	5,465

Table 6.3: Genre distribution in Million Song Dataset.

row genres such as "Animation", "Documentary", "Film-Noir" or "Western" are present in a relatively smaller number of movies than more general genres such as "Drama", "Comedy" or "Thriller". Genres do not form disjoint categories, as seen in Figure 6.1. One can see that, for instance, there are only 46 pure "Romance" movies, and the other 92% of movies in this genre overlap with at least one other genre. Other genres also have a high degree of overlap. In fact, there is no clear hierarchical structure between the genres. It also seems that overlaps between genres do not follow any particular distribution. Furthermore, pairwise overlaps between genres are not wide enough as to establish any clear sub-genre relationship between one another; even the narrowest and most specific genres (for example, Crime) have only partial overlaps (<60%) with more general genres such as Drama. Table 6.2 shows the corresponding genre distribution for the Netflix dataset. We omit the corresponding figure showing the overlap in this dataset as it is equivalent – saving the higher number of movies and genres – to that of MovieLens1M. This provides a confirmation of the generality of the previous assertions in the movie recommendation scenario.

Table 6.3 and Figure 6.2 provide the observations for genre generality and overlap in the music recommendation setting of the Million Song Dataset. In this case, even when the genres were chosen to avoid too general genres and promote a balance in the generality of each (refer to Chapter 3 for details), we still can see that the number of songs in each genre varies significantly, from 28,246 songs in the more general "Alternative Rock" to 5,354 in narrower "Reggae". The overlap of genres is also observed in this domain in a much smaller but noticeable degree. For example, 36% of "Alternative Rock" songs are also described with other genres and a more specific genre such as "Dance" frequently co-occurs with other genres such as "Jazz" or "Alternative Rock".

Finally, the information regarding the distribution of the number of covered genres by each item is shown in Table 6.4 for our three datasets and the TREC Web Tracks from 2009 to 2012. We treated separately the documents in TREC according

	1	2	3	⩾4
TREC-a	82.1%	13.3%	3.5%	1.1%
TREC-u	57.9%	27.7%	10.3%	4.1%
ML1M	38.9%	34.9%	17.9%	8.3%
Netflix	22.3%	33.1%	25.8%	18.9%
MSD	60.5%	22.9%	10.2%	6.4%

Table 6.4: Distribution of the number of genres covered by each item in the TREC Web Tracks from 2009 to 2012 and MovieLens1M, Netflix and Million Song Dataset.

to whether they are relevant to ambiguous (TREC-a) or underspecified (TREC-u) queries. As we can observe, more than 80% of the documents in the TREC relevance judgments for ambiguous queries cover only one subtopic, which greatly contrasts with the movie recommendation datasets, where most of the movies cover more than one genre. The case of the Netflix dataset is specially revealing: almost 19% contain a really high overlap of four or more genres. As a midpoint, the distribution of the number of overlaps in TREC documents for underspecified queries and music genres in the Million Song Dataset presents a smaller but noteworthy number of overlaps: in both cases, around 40% of the documents/songs cover more than one subtopic/genre.

6.4 measuring genre diversity in recommendation lists

We all have an intuitive idea of what genre diversity means for a list of movies or songs. Yet when it comes to translating the intuition to a mathematical expression that reflects degrees of diversity by a numeric value, one has to be more specific about what the value should reflect. In particular, we draw from the literature about diversity in Recommender Systems (Ziegler et al., 2005; Zhang and Hurley, 2008) and Information Retrieval (Agrawal et al., 2009; Carbonell and Goldstein, 1998; Zhai et al., 2003) and our contributions in the previous chapters to determine the two different dimensions that should be considered to this respect, namely genre coverage and redundancy. We take them as required properties that a genre-based recommendation diversity metric should capture. Furthermore, we argue that these dimensions should be captured in a way that takes into account the properties of genres discussed and exemplified in Section 6.3. Moreover, we add to these a third and new requirement, size-awareness, which has not been explicitly considered in prior work. We discuss each of these three properties next.

Coverage is the simplest and most obvious property. Since most users enjoy items from a variety of genres, it is important that the recommendation list covers as many of them as possible. Coverage relates to the *subtopic retrieval* problem

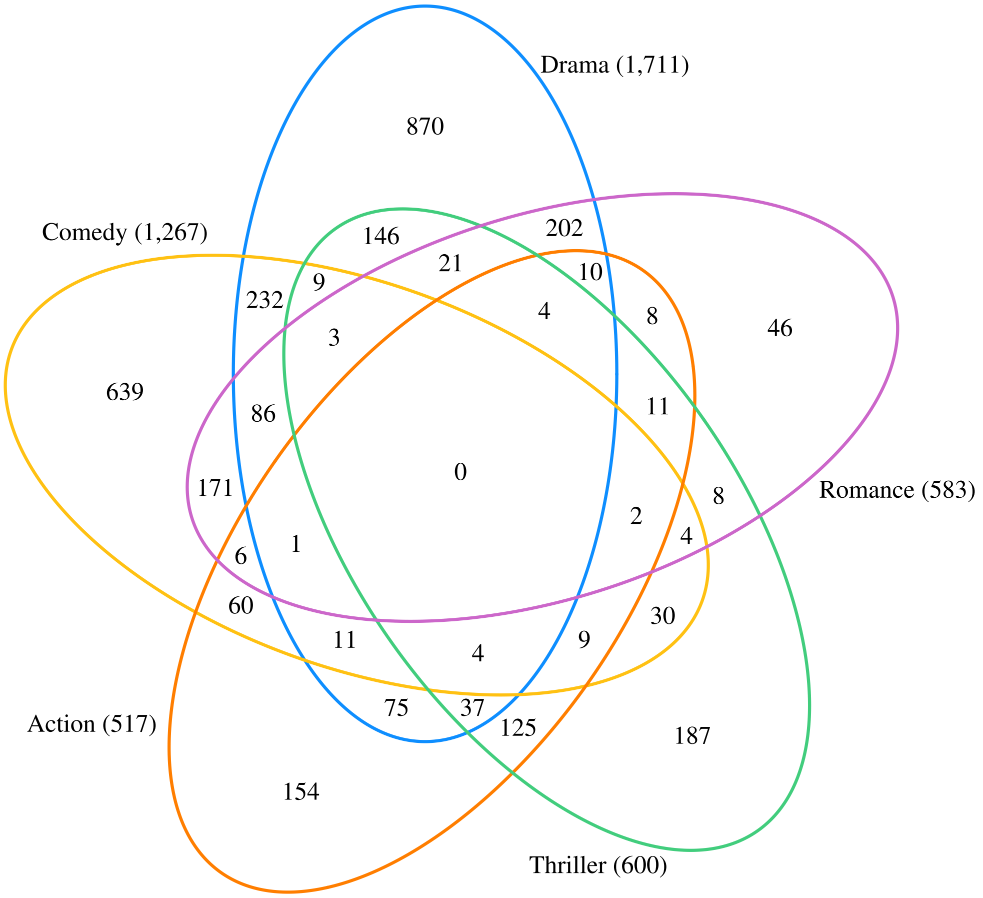

Figure 6.1: Venn diagram for the 5 most frequent genres in the MovieLens1M dataset.

of Zhai et al. (2003) and, more generally, to the Intent-Aware framework for evaluating search result diversification (Agrawal et al., 2009; Clarke et al., 2008; Santos et al., 2010a) and our adaptation of it to the recommendation problem in Chapter 5. Moreover, this coverage should be proportional: even when a user is interested in several genres, the personalized importance of each genre is not equal. Therefore, the more a user is interested in a given genre, the more important it is that the genre is covered in the recommendation list. The idea of proportional coverage appears in the Proportionality framework of Dang and Croft (2012).

Second, **redundancy** should also be considered. It is not enough to have a high coverage of genres in order to have a diverse recommendation list. We may put it this way: it is as important to present items that cover a certain genre as to present other items that do *not* cover it. This notion of redundancy should take into account the preferences for the user as well as how general each genre is. Consider the extreme example shown in Table 6.5 where three movies are recommended to

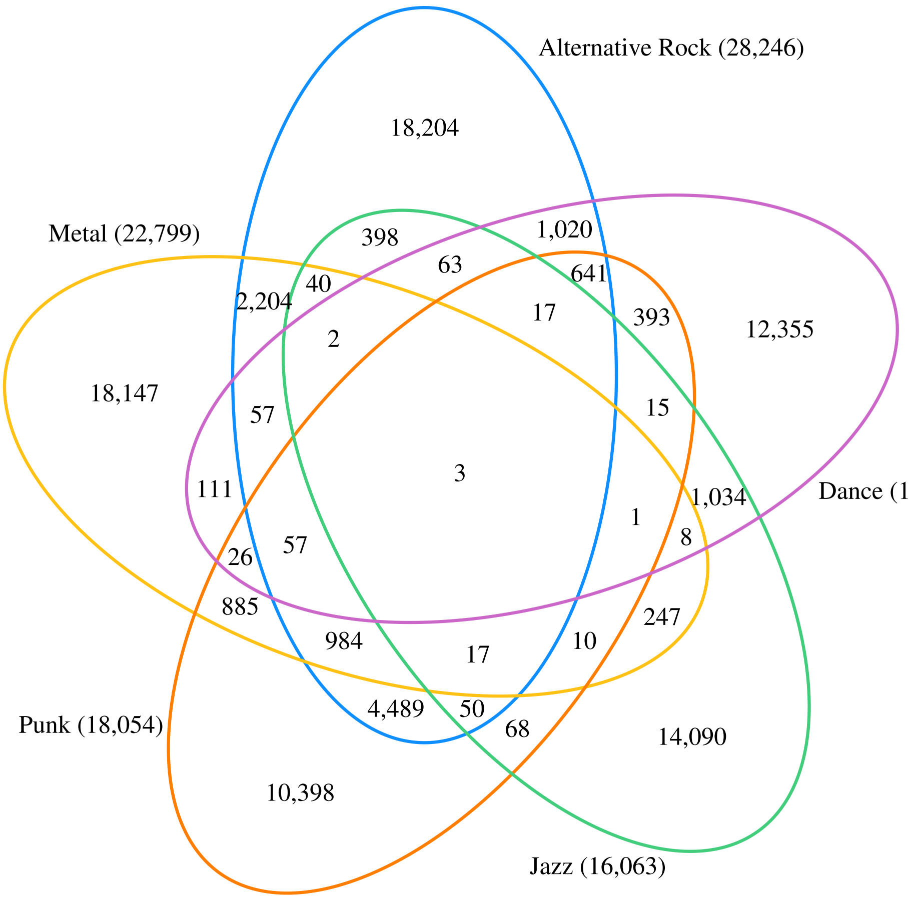

Figure 6.2: Venn diagram for the 5 most frequent genres in the Million Song Dataset.

a user. Even if these 3 movies cover a total of 6 genres, the diversity is not quite perceivable. This is because all three movies cover a very narrow "Western" genre which makes the recommendation list highly redundant. To some extent, redundancy is regarded in prior work in diversity in Recommender Systems (Smyth and McClave, 2001) and Information Retrieval (Carbonell and Goldstein, 1998) in the form of minimizing pairs of similar (redundant) items or documents, respectively. Although most of the proposals for evaluation and enhancement in the Intent-Aware framework also consider some notion of redundancy, we see later that this is posited in a different manner as intended in our proposal.

Finally, **size-awareness** is taken into account. Coverage and redundancy should depend on the length of the recommendation list. Since the rise of mobile devices, the issue of having limited screen real estate to show recommendations requires a careful selection of what to display in that list. We also improve over existing diversity enhancing techniques by specifically addressing the recommendation list size. For example, when generating a short recommendation list one should only recommend items from the most relevant genres. In a longer list we could have higher genre redundancy depending on the generality of the involved genres. List size-awareness is considered implicitly in the work of of Dang and Croft (2012) by considering the desired proportion of documents covering a certain aspect in a search result list. In our work, we elaborate on the notion of size-awareness by making it play a central role on the assessment and enhancement of coverage and redundancy. To the best of our knowledge this kind of adaptation has not been explored in prior work on search or recommendation diversity.

The reviewed techniques in Section 6.2 do not satisfy all these properties, in particular:

- The Intra-List Distance of Smyth and McClave (2001) and our generalized, equivalent Expected Intra-List Distance metric of Chapter 4 are defined as a pairwise property of elements in a list. A pair-wise property does not translate however as directly as we may expect to a list-wise property as we are stating. Further, it is not trivial to consider a similarity measure that takes into account by itself the generality of different genres (are two dramas as similar as two westerns?) and the user-specific importance of each of them.

- The Intent-Aware framework (IA-metrics, IA-select and xQuAD) (Agrawal et al., 2009; Santos et al., 2010a) considers coverage and a concept of redundancy, but as to the latter, the scheme does not fully capture the view that it is equally important to present items that cover a certain genre as to present other items that do not cover it. Specifically, the redundancy component of ERR-IA and xQuAD reduces the contribution of items that cover redundant genres, rather than discounting them as negative from the list diversity value. Thus, items covering a redundant genre will contribute positively to the diversity even though the contribution diminishes with each additional occurrence of the genre. Furthermore, this redundancy does not detract at all from the contribution of additional genres the items can have in addition to the redundant one – that is, the genres are assumed to be totally independent from each other. This effect is aggravated in the cases where we have genres highly overlapping between each other, as in the case of movie recommendations. The example in Table 6.5 illustrates this effect: it is fine (diversity-wise) in the context of this framework that all the movies in the recommendation list be westerns, as long as they cover also other genres. As a consequence, the diversifications are biased to retrieve items that cover many genres. We may reasonably question the implicit assumption in this scheme that multiple genres in the same item will procure the same diversity perception as multiple genres over different items.

Movie	**Genres**
Wild Wild West	Action, Comedy, Sci-Fi, Western
Cowboys and Aliens	Action, Sci-Fi, Thriller, Western
The Good, the Bad and the Ugly	Adventure, Western

Table 6.5: Example of redundant movie recommendations.

- The work by Dang and Croft (2012) does cover an idea of user-centric proportionality, but over-representation – and thus, redundancy – is not penalized and therefore, it may also suffer from the same problems as xQuAD for genre diversity.
- None of the prior search or recommendation diversification methods takes into account the size of the retrieved list that will be presented to (or browsed by) the user. The diversification schemes have therefore no means to consider this information to enhance diversity at a particular list size.

These issues indicate that the previous approaches, even when thy provide sound and effective solutions to assess and enhance the diversity of recommendations, may result sub-optimal when relying on genres as the source of diversity as they do not properly address all our proposed requirements. The case of the Intent-Aware approaches is specially relevant since it establishes a clear difference between subtopics and genres: in the former, redundancy is used to maximize the number of documents covering different subtopics as early as possible in the ranking, while in the latter, genre redundancy is used to minimize the number of items covering a genre. This difference is even more crucial in recommendation domains where genres overlap highly with each other.

In the next section, we propose a new framework to adequately fulfill coverage, redundancy and size-awareness and thus avoid the pitfalls manifested in the previous approaches.

6.5 a binomial framework for genre diversity

A naïve approach for creating diverse recommendations consists in making a random selection of items. This approach offers highly diverse recommendations – as seen in Chapter 4 –, but it tends to approximate the poorest possible output in terms of the relevance of recommendations for the user interests, which makes it an option of little practical use. Still, the nature of the selection of genres in a random recommendation provides a meaningful basis to build a revised notion of diversity upon it. In particular, we propose to use a binomial distribution to model how a personalized recommendation would match a random recommendation in terms of the diversity of genres, using the binomial distribution to model the likelihood that a given genre will appear by chance in a recommendation, and take this as a reference to assess the diversity value of a given genre distribution among recommended items. In essence, this approach means considering random item recommendation as the optimal approach in terms of pure genre diversity, and using a binomial distribution as the model for the genre distribution resulting from random item sampling.

6.5.1 *The Binomial Diversity Metric*

The binomial distribution is the discrete probability distribution of the number k of successes in a sequence of N independent Bernoulli trials with the same probability of success p. A random variable that follows this distribution, $X \sim B(N, p)$, has the following probability mass function:

$$P(X = k) = \binom{N}{k} p^k (1-p)^{N-k} \tag{6.1}$$

We base our definition of a genre diversity metric on top of this as follows. For each genre, we measure its coverage and redundancy using binomial distributions. We consider the selection of an item covering each genre as a Bernoulli trial, whereby for each genre, a recommendation list can be viewed as a sequence of Bernoulli trials. It must be noticed that these trials are not independent: a recommendation list is actually a selection without replacement. However, given that the typical recommendation list size is usually much smaller than the set of movies covering each genre, we can treat these trials as if they were independent, and therefore use the binomial distribution to model how likely is a genre to appear in a recommendation list. See Table 6.6 for a mapping between the probabilistic notation and the natural item/genre terminology.

distribution		recommendation
trials	N	recommendation list size
probability of success	p_g	probability of a genre
no. of successes	k_g	no. of items covering a genre

Table 6.6: Binomial distribution for genre diversity.

More formally, for an item i and the set of genres $\mathcal{G}_i$ it covers, we consider the experiment of a randomly sampled genre g belonging to $\mathcal{G}_i$. Given a genre g and a recommendation list R of size N, we denote the number of recommended items belonging to that genre, that is, the number of "successes", as:

$$k_g = |\{i \in R \,:\, g \in \mathcal{G}_i\}|$$

We take the probability of a genre p_g as a measure of how likely is the number k_g of items covering a genre g in that recommendation. As required in Section 6.4, this probability should take into account the generality of a genre and also the relevance of each genre for the user. We propose to combine global genre distribution statistics and personalized user preferences to estimate p_g as follows.

On the one hand, the relevance of a genre for the user u can be estimated by using historical data, i.e., considering the local proportion p_g^u of the items the user has had some interaction with, denoted as $\mathcal{I}_u$. On the other hand, the generality of the genre can be estimated by the global proportion $p_g^{\mathcal{U}}$ of items in the user preferences covering it. To join both global and local probabilities, we propose a simple linear combination:

$$\begin{aligned} p_g^u &= \frac{|\{i \in \mathcal{I}_u \,:\, g \in \mathcal{G}_i\}|}{|\mathcal{I}_u|} \\ p_g^{\mathcal{U}} &= \frac{\sum_v |\{i \in \mathcal{I}_v \,:\, g \in \mathcal{G}_i\}|}{\sum_v |\mathcal{I}_v|} \\ p_g &= (1-\alpha)\, p_g^{\mathcal{U}} + \alpha\, p_g^u \end{aligned} \tag{6.2}$$

With all the components of genre-based binomial distributions, we now define scores for the coverage and redundancy of a recommendation list R. We measure coverage as a property defined by the genres that are present in the recommendation list and those that are not. The maximum coverage would be achieved when all the genres of interest are covered in the recommendation list. However, this maximum is not always reachable, especially in small recommendation lists. Therefore, when some genres cannot be covered, the coverage should reflect the loss caused by their absence, which should be proportional to their importance. Moreover, the importance, and so the potential loss, may vary significantly between genres, so we propose an aggregation of the coverage scores for each genre in the form of the geometric mean. We thus define the **Binomial Coverage** score as the product of the genres not represented in the recommendation list of their probabilities of not being randomly selected according to X_g, normalized by the $|\mathcal{G}|$-th root:

$$\text{Binom-Cov}(R) = \prod_{g \notin \mathcal{G}_R} P(X_g = 0)^{1/|\mathcal{G}|} \tag{6.3}$$

We define redundancy, in turn, only by the genres covered in the recommendation list. The moment one genre appears more than once in a recommendation list, it can be potentially redundant, although not all genres will be equally affected. We model the redundancy of a genre appearing k_g times in a recommendation list

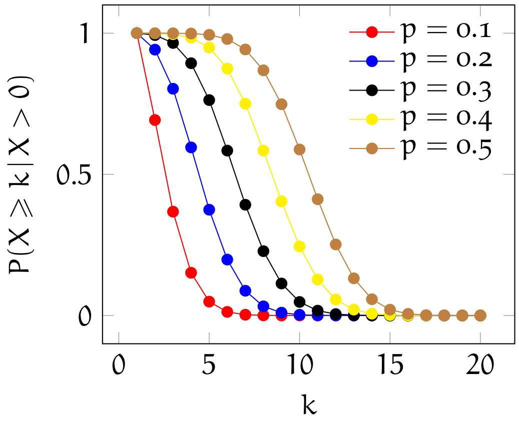

Figure 6.3: $P(X \geqslant k\,|\,X > 0)$ for different values of p and k of binomial distributions with $N = 20$ (continuous lines are drawn just as a reference).

by a "remaining tolerance" score that reflects how probable it would be that the genre appeared at least k_g times in a random list:

$$P(X_g \geqslant k_g\,|\,X_g > 0) = 1 - \sum_{l=1}^{k_g-1} P(X_g = l\,|\,X_g > 0) \tag{6.4}$$

Some examples of this "remaining tolerance" score are illustrated in Figure 6.3. Again, we summarize the redundancy penalization for all genres by means of the geometric mean, in this case restricted to the genres present in the recommendation list – uncovered genres are, by definition, non redundant. The **Binomial Redundancy** score is consequently defined as the product of the "remaining tolerance" scores for each covered genre, normalized by the $|\mathcal{G}_R|$-th root:

$$\text{Binom-Red}(R) = \prod_{g \in \mathcal{G}_R} P(X_g \geqslant k_g\,|\,X_g > 0)^{1/|\mathcal{G}_R|} \tag{6.5}$$

The **Binomial Diversity** metric is then defined as the product of both components:

$$\text{Binom-Div}(R) = \text{Binom-Cov}(R) \cdot \text{Binom-Red}(R) \tag{6.6}$$

The previous definition can be adapted to consider only the relevant recommended items by re-defining k_g as the number of relevant items covering the genre g and the number of trials N as the number of relevant recommended items.

Note that the Binomial Diversity satisfies all the properties described in Section 6.4. It maximizes the coverage of the genres according to their p_g. It takes into account user preferences via p_g^u. It penalizes over-represented genres by rapidly decreasing their redundancy score. Lastly, it is adapted to the recommendation length by parameter N.

6.5.2 *Binomial Re-ranking Strategies*

Following the idea of greedy re-ranking strategies to optimize Intra-List Diversity – present in the literature of Information Retrieval (Agrawal et al., 2009; Carbonell and Goldstein, 1998; Dang and Croft, 2012; Santos et al., 2010a) and Recommender Systems (Ziegler et al., 2005; Zhang et al., 2012) and exploited in previous chapters –, we now propose the corresponding re-ranking strategies for Binomial Coverage, Redundancy and Diversity, which are straightforwardly derived from the proposed metric scheme and the greedy re-ranking scheme firstly introduced in Section 4.8. Concretely, we consider the re-rankings defined by taking as the novelty component $nov(i \mid S)$ the score provided by the relevance-unaware target metrics when the candidate item i is added to the previously re-ranked items S:

$$nov_{\text{Binom-Cov}}(i \mid S) = \text{Binom-Cov}(S \cup \{i\}) \tag{6.7}$$

$$nov_{\text{Binom-Red}}(i \mid S) = \text{Binom-Red}(S \cup \{i\}) \tag{6.8}$$

$$nov_{\text{Binom-Div}}(i \mid S) = \text{Binom-Div}(S \cup \{i\}) \tag{6.9}$$

Note that, as opposed to prior re-ranking strategies, our Binomial re-rankings consider explicitly in the novelty component the list size N of the final re-ranked recommendation list as it is one integral parameter of the metric. This introduces a new and original perspective for the optimization of a particular recommendation size list which improves over the standard greedy approaches that, implicitly, assume an unbounded re-ranking of a recommendation list.

6.5.3 *Qualitative analysis*

In addition to the empirical behavior of the proposed scheme, the Binomial Diversity metric fulfills qualitative properties that further specify the requirements stated earlier in Section 6.4. These properties can be formalized by four postulates shown in Table 6.7, which we propose as a basis on which diversification metrics can be analyzed and compared to each other, providing a clear way to show properties of each metric, identify and report the differences, in a similar perspective as proposed in (Amigó et al., 2013). Each postulate presents a rule, which expresses a simple idea on how we can reason about the genre-based diversity. We represent each of the postulates by providing two ranked lists of items (displayed horizontally in the table) with minimal differences. The ranked list denoted by "Better" should have strictly higher diversity that the one denoted by "Worse". For example, the first postulate expresses the idea that a ranked list of two items that cover two genres (a and b) is more diverse than a list of two items that cover only one genre (a). We mark a method with "Yes" only if the metric complies with the postulate, otherwise we indicate to what extent the metric fails to satisfy the postulated inequality (either the metric yields the opposite inequality, or is insensitive to the difference between the two lists). We can see that all of the state of the art methods fail at least one of the tests, and only our proposed Binomial Diversity that combines Binomial Coverage and Redundancy properties complies with all the postulates. For illustration, we show in the same table the diversity score that each of the analyzed diversification metric assigns to the prototypical lists.

In order to further illustrate how the diversification metric works and to show the benefits of the genre-based approach, we may examine the working example shown in Table 6.8. The example shows the top 20 recommended movies by the item-based nearest neighbors (IB) method ($R_0 \cup R_1$) for a sample user from the Netflix dataset, and the re-ranking of this list by the binomial diversification ($R_0 \cup R_2$), shown by the movies that are removed (R_1) and added (R_2) as a result of the re-ranking. The first row of the table summarizes the user taste profile (p_g^u), i.e. what fraction of movies of each genre he has rated. We see that the user is inclined towards Drama, Comedy and Action movies. We may also notice that the user seldom watched War movies. Both recommendation lists have an overlap of 11 movies (R_0) that are shown below the user profile information. If we compare the differences between both recommendation lists – the IB baseline $R_0 \cup R_1$ and its diversification by the binomial scheme $R_0 \cup R_2$, we see that the baseline promotes Action and War movies that are over-represented in the final list of 20 movies, thus creating a highly redundant recommendation. The recommender under-represents other genres such as Comedy which plays a major part in the user profile. The binomial diversification, on the other hand, uses the p_g^u and the list size as the reference for how many movies of each genre it should select to avoid redundancy. There are already 7 Action movies in the list and, therefore, it promotes several Comedies instead. Moreover, it includes new genres such as Animation, Children's and Mystery that help improve the coverage score. This leads to a significant increase of the diversification score for the diversified list.

6.6 EXPERIMENTS

In order to show the properties of the Binomial Diversity framework, we have carried out a series of experiments in the context of the experimental design of Chapter 3. In particular, we focus on analyzing the following aspects:

- The results of Binomial Coverage, Redundancy and Diversity on the baseline recommendation algorithms on the three considered datasets and their comparison with metrics from the rest of the compared approaches in Section 6.2.

	i_1	i_2		**ILD**	**S-recall**	**ERR-IA**	**CPR**	**Cov**	**Red**	**Div**
				Yes	Yes	**Not Always**	**Not Always**	Yes	Yes	Yes
Better	b	a	Regardless of	1.0000	0.6666	**0.4000**	**0.7857**	0.8255	1.0000	0.8255
Worse	a	a	p(a), p(b)	0.0000	0.3333	**0.5000**	**0.8571**	0.6814	0.3333	0.2269
				No(=)	Yes	Yes	Yes	Yes	**No(=)**	Yes
Better	a c	b	Regardless of	**1.0000**	1.0000	0.7000	0.9643	1.0000	**1.0000**	1.0000
Worse	a	b	p(a), p(b), p(c)	**1.0000**	0.6666	0.5000	0.8929	0.8255	**1.0000**	0.8255
				Yes	**No(=)**	**No(<)**	**No(<)**	**No(=)**	Yes	Yes
Better	a	b	Regardless of	1.0000	**0.6666**	**0.5000**	**0.8929**	**0.8255**	1.0000	0.8255
Worse	a b	b	p(a), p(b)	0.5000	**0.6666**	**0.6500**	**0.9286**	**0.8255**	0.3780	0.3120
				No(=)	**No(=)**	Yes	Yes	Yes	Yes	Yes
Better	a	a	p(a) > p(b)	**0.0000**	**0.3333**	0.5000	0.8571	0.6814	0.3333	0.2271
Worse	b	b		**0.0000**	**0.3333**	0.2500	0.6429	0.5200	0.1429	0.0743

Table 6.7: Postulates of genre-based diversity. Each of the four postulates shows two rankings (displayed horizontally) with better or worse diversity. Each item in the ranking is represented by the genres (a, b, c) that it belongs to. We show in the table the diversity score that each of the metric assigns to the lists. In the computation of the metric values, we assume for simplicity there are only three genres in the dataset, with prior probabilities, as an example, p(a) = 0.5, p(b) = 0.25, p(c) = 0.25. For ERR-IA we use the same definition as in the TREC diversity task (as computed by the ndeval script), generalized to support non-uniform aspect distributions.

	Movie	Action	Adventure	Animation	Children's	Comedy	Crime	Drama	Mystery	Romance	Sci-Fi	Thriller	War	Western
	$\mathbf{p_g^u}$	0.25	0.20	0.04	0.08	0.39	0.13	0.42	0.07	0.19	0.15	0.25	0.06	0.02
Kept (R_0)	Braveheart	X						X					X	
	Jerry Maguire							X		X				
	Matrix, The	X									X	X		
	Negotiator, The	X										X		
	Patriot Games	X										X		
	Pulp Fiction						X	X						
	The Silence of the Lambs							X				X		
	Terminator 2	X									X	X		
	Titanic							X		X				
	Total Recall	X	X								X	X		
	True Lies	X	X			X				X				
Removed (R_1)	Air Force One	X										X		
	Enemy of the State	X										X		
	Get Shorty	X				X		X						
	Gladiator	X						X						
	Green Mile, The							X				X		
	Independence Day	X									X		X	
	Schindler's List							X					X	
	Star Wars: Episode V	X	X					X			X		X	
	Star Wars: Episode VI	X	X							X	X		X	
Added (R_2)	As Good As It Gets					X		X						
	Back to the Future III					X					X			X
	Elizabeth							X						
	Erin Brockovich							X						
	The Game								X			X		
	Leon: The Professional						X	X		X		X		
	South Park			X		X								
	There's Sth. About Mary					X								
	Toy Story			X	X	X								

Table 6.8: Binomial diversification in action.

- The effect or re-ranking diversification techniques, specially the interplay between the metrics and diversification methods of the different frameworks for assessment and enhancement of Intra-List Diversity.

- The analysis of the specific properties of our framework: the balance between genre generality and user preferences (controlled by the α parameter in Equation 6.2) and the size-awareness of the framework.

6.6.1 *Evaluation of Baseline Recommendation Algorithms*

We measure the performance of the baseline recommendation algorithms in the three datasets of our experimental design with the metrics of our Binomial framework, namely Coverage, Redundancy and Diversity at cut-off 20 and the genre distribution parameter to $\alpha = 0.5$ to show a balance between generality and user relevance of the genres. Additionally, we also measured these baseline recommendations with metrics from the different frameworks using genres: the cosine-based, rank-unaware EILD metric, the CPR metric of the Proportionality framework and S-recall and ERR-IA representing the Intent-Aware metrics. All considered metrics, except S-recall and ERR-IA, were tested in both relevance-unaware and aware configurations.

Table 6.9 shows the results of the relevance-unaware metrics and Table 6.10 does the corresponding for the relevance-aware metrics. The results illustrate the properties of our metrics and shows the relations between our framework and the compared ones. In general, we can observe that our Binomial Coverage, as expected, shows a high degree of agreement with S-recall and the Redundancy score is somewhat similar to EILD. The combination of these scores, which constitutes our Binomial Diversity metric, seems to be mostly dominated by its Redundancy component, although the Coverage introduces considerable nuances. We next comment the particular results for each dataset and configuration.

The relevance-unaware results of our framework in Table 6.9 mainly manifest that, in the absence of relevance, the random recommendation is the most diverse recommendation due to his low redundancy. This result is expected, since our Binomial framework is inspired by the nature of genre-selection in random recommendations. In the MovieLens1M dataset, we can see though that the random recommendation has the worst coverage while the popular recommendation has the best coverage at the cost of being highly redundant, which affects negatively the combined diversity. Regarding the personalized algorithms, they all have a similar coverage scores although UB is slightly better than the rest. In terms of redundancy, we see that iMF stands out as the least redundant of the personalized recommendations – although not as good as the random recommendation. When coverage and redundancy scores are combined, the iMF offers the most diverse recommendations, followed by UB given its high coverage. The results in the Netflix dataset present some differences with respect to MovieLens1M, specially in the performance of the random recommendations and the differences between latent factors and nearest neighbors algorithms. Concretely, in this dataset the random recommendation achieves also the highest coverage scores, while the popular recommendations are the worst choice in terms of coverage and redundancy. Considering the personalized recommendations, we clearly observe a distinction between the two families of recommendation algorithms for this dataset: nearest

		Cov	Red	Div	EILD	CPR	S-recall
ML1M	**Rnd**	0.7184	0.6140	0.4448	0.7616	0.6893	0.6691
	Pop	0.8392	0.2321	0.1973	0.7010	0.7099	0.7206
	iMF	0.7858	0.3582	0.2875	0.6434	0.8227	0.6621
	pLSA	0.7728	0.3217	0.2558	0.6320	0.8051	0.6542
	UB	0.8189	0.3292	0.2761	0.6704	0.8158	0.6872
	IB	0.7923	0.3211	0.2591	0.6725	0.8103	0.6445
Netflix	**Rnd**	0.8541	0.4139	0.3548	0.7885	0.7635	0.6025
	Pop	0.8051	0.1767	0.1437	0.7042	0.8300	0.4091
	iMF	0.8057	0.2433	0.2004	0.6515	0.7855	0.4861
	pLSA	0.8131	0.2578	0.2140	0.6596	0.7979	0.4756
	UB	0.8393	0.2806	0.2382	0.6843	0.8209	0.4888
	IB	0.8410	0.2774	0.2355	0.7003	0.8278	0.4739
MSD	**Rnd**	0.7429	0.4264	0.3191	0.9070	0.5257	0.7507
	Pop	0.8163	0.1409	0.1149	0.7930	0.6582	0.7701
	UB	0.7127	0.1418	0.1106	0.6455	0.7109	0.6107
	IB	0.6532	0.1199	0.0858	0.6169	0.6644	0.5285

Table 6.9: Results of relevance-unaware metrics for the baseline recommendation algorithms in MovieLens1M, Netflix and Million Song Dataset.

neighbors are consistently better at coverage, redundancy and diversity than latent factor approaches. Finally, in the music recommendation setting of the Million Song Dataset, we see interesting differences with respect to the previous datasets. In particular, we see that one of the personalized recommendations, the item-based nearest neighbors, is the worst choice according to all criteria, while the user-based variant is as good as the most-popular recommendations in terms of redundancy and final diversity.

Regarding the relevance-aware results of our framework in Table 6.10, we clearly observe how random and popularity, which showed some competitiveness – specially the first one – when relevance was not considered, are now the worst choices due to their low relevance compared to personalized recommendations. Regarding the personalized algorithms, the results of our framework in the presence of relevance present some changes. In the MovieLens1M dataset, we see now that UB is the algorithm that offers the least redundant recommendations, and therefore is the second best choice after iMF for Binomial Diversity. In a completely different direction, pLSA seems to provide highly redundant relevant recommendations, and despite its high accuracy it is a particularly bad choice in Binomial Diversity.

		Cov	Red	Div	EILD	CPR	ERR-IA
ML1M	Rnd	0.0527	0.0979	0.0126	0.1551	0.0099	0.0044
	Pop	0.2433	0.5224	0.1435	0.2060	0.2103	0.0962
	iMF	0.3585	0.5486	0.1906	0.2408	0.4090	0.2212
	pLSA	0.3302	0.5398	0.1746	0.2253	0.3669	0.1918
	UB	0.3407	0.5562	0.1858	0.2343	0.3739	0.2049
	IB	0.3163	0.5495	0.1744	0.2251	0.3371	0.1832
Netlfix	Rnd	0.1602	0.0404	0.0111	0.1588	0.0039	0.0014
	Pop	0.3148	0.4234	0.1606	0.1857	0.1586	0.0670
	iMF	0.4158	0.4955	0.2161	0.2200	0.3041	0.1412
	pLSA	0.4188	0.4768	0.2130	0.2222	0.3137	0.1429
	UB	0.4391	0.4928	0.2275	0.2327	0.3356	0.1562
	IB	0.4101	0.5006	0.2208	0.2200	0.2785	0.1250
MSD	Rnd	0.1919	0.0012	0.0003	0.1814	0.0001	0.0000
	Pop	0.2230	0.2073	0.0638	0.1703	0.0434	0.0098
	UB	0.3300	0.4415	0.1566	0.1735	0.2510	0.1034
	IB	0.3312	0.4338	0.1536	0.1724	0.2480	0.0959

Table 6.10: Results of relevance-aware metrics for the baseline recommendation algorithms in MovieLens1M, Netflix and Million Song Dataset.

The same under-performance of pLSA in terms of relevance-aware redundancy is also present in the Netflix dataset. Finally, the results of the Binomial framework in the Million Song dataset reveal that, when considering the coverage and redundancy of the personalized recommendations, both algorithms show almost indistinguishable outcomes.

Concerning the relationship between the metrics of our framework and other compared approaches, we can observe in Table 6.9 and the following similarities and differences. First, the results of S-recall tend to resemble those of our our Binomial Coverage score in all three datasets. Second, the results of CPR mostly disagree with our metrics, in particular it contradicts our Binomial Redundancy and Diversity by assigning the worst diversity scores to the random recommendation. Third, the EILD metric, which captures the redundancy between pairs of items, seems to be related to our redundancy score, as the results for the Netflix dataset exemplify. In the results for relevance-aware metrics in Table 6.10 we see that relevance mainly dominates the results and thus it mostly fades the trends observed in the relevance-unaware case. However, the behavior of the Binomial Redundancy differs from most of the metrics when considering relevance. In the next part of

	P	Cov	Red	Div	EILD	CPR	S-recall	GPI
iMF	0.2335	0.7858	0.3582	0.2875	0.6434	0.8227	0.6621	2.7010
+Cov (1.0)	0.2141	0.9788	0.4313	0.4227	0.6968	0.9395	0.8823	3.0397
+Red (1.0)	0.1386	0.9192	0.9286	0.8536	0.7643	0.8174	0.7981	1.7158
+Div (1.0)	0.1174	0.9717	0.8803	0.8557	0.7409	0.9244	0.8580	1.9913
+ILD (1.0)	0.1551	0.9252	0.5237	0.4922	0.8195	0.8003	0.8157	2.3788
+PM (1.0)	0.2218	0.8770	0.3952	0.3492	0.6700	0.8978	0.7135	2.8801
+xQuAD (1.0)	0.1818	0.9313	0.1853	0.1739	0.6400	0.9105	0.7865	3.7615

Table 6.11: Results of relevance-unaware metrics for the re-rankings in MovieLens1M.

	P	Cov	Red	Div	EILD	CPR	ERR-IA	GPI
iMF	0.2335	0.3585	0.5486	0.1906	0.2408	0.4090	0.2212	2.7010
+Cov (0.3)	0.2283	0.3663	0.5643	0.2022	0.2468	0.4057	0.2327	2.8014
+Red (1.0)	0.1386	0.2093	0.7073	0.1711	0.2296	0.2605	0.1744	1.7158
+Div (0.5)	0.2179	0.3517	0.6133	0.2184	0.2510	0.3778	0.2143	2.5273
+ILD (0.6)	0.1927	0.3138	0.6358	0.2036	0.2690	0.3371	0.2003	2.4403
+PM (1.0)	0.2218	0.3773	0.5633	0.2072	0.2428	0.4035	0.2200	2.8801
+xQuAD (0.4)	0.2301	0.3860	0.5381	0.1981	0.2413	0.4119	0.2334	3.0678

Table 6.12: Results of relevance-aware metrics for the re-rankings in MovieLens1M.

the experimental results, we analyze in depth the relation between frameworks by testing the effects of re-ranking strategies from one framework when evaluated with metrics from another.

6.6.2 *Evaluation of Re-Ranking Strategies*

We now analyze the effect of the re-ranking strategies detailed in Section 6.5.2 and those of the compared approaches when applied to the personalized recommendations in MovieLens1M. This evaluation not only shows how Binomial re-ranking strategies can enhance the diversity as measured by our Binomial framework, but also provide further details about how the different diversity frameworks are related to each other. Specifically, we applied our three re-ranking strategies, the partial coverage and redundancy diversifications and the Binomial Diversity, and the ILD, PM and xQuAD diversifications targeting EILD, CPR and S-recall/ERR-IA, respectively. Additionally, we measure the precision and the average number of genres per recommended item (GPI).

In Tables 6.11 and 6.12 we show the results of applying re-ranking strategies in terms of the different metrics of the analyzed frameworks in their relevance-unaware and aware variants. For each diversifier, we selected the value of the λ relevance-diversity trade-off parameter that achieves the best results with respect to its target metric. Additionally, in Figures 6.4 and 6.5 we show the same results in terms of the relative improvement with respect to the original ($\lambda = 0.0$) recommendation. A first observation is that all the diversifications involve a decrease in the accuracy of the recommendations as measured by precision, showing an also expectable trade-off between relevance and diversity. A second and expected observation is that, for almost every metric, its corresponding re-ranking diversifier achieves the best performance, specially our Binomial re-ranking diversifications. A third observation comprises the results of the relevance-aware CPR, for which none of the diversifications, including PM, are able to show significant improvements. This shows that this diversification framework, originally devised for a search task, may not get the expected results in a recommendation setting with a different subtopic-document (in our case, genre-item) distribution pattern, which is one of the motivations for our framework. The remaining observations address the analysis of the interaction between diversifications and metrics of the different frameworks.

As we can see, the partial Binomial Coverage re-ranking is particularly helpful at enhancing relevance-unaware CPR, S-recall and ERR-IA, showing that both the Proportionality and Intent-Aware frameworks are driven by genre coverage. Respectively, the PM and xQuAD diversifications also seem to be effective and optimizing Binomial Coverage. The Binomial Redundancy re-ranker, in turn, only seems to improve the relevance-unaware version of EILD and S-recall. We think this behavior is mainly caused by the marked decrease in precision of this diversification. The joint Binomial Diversification, however, seems to balance appropriately its two components and achieves good improvements in terms of EILD, CPR and S-recall without relevance, but, adding relevance, only a minor improvement in EILD and a drop in performance for CPR and and ERR-IA. The correspondence between our redundancy score and the Pair-Wise framework is observed in the results of the ILD diversification, which consistently improves the scores of the Binomial Redundancy and Diversity in all their variants. Finally, the weaknesses of the Intent-Aware framework are manifested by the results of the xQuAD in terms of our Binomial Redundancy and Diversity and the GPI score. In particular, we see that xQuAD creates highly redundant recommendations are measured by our relevance-unaware metrics. Also, the results of the GPI score reveal that xQuAD is biased toward selecting items covering many genres so that, even when favoring a high coverage, causes the recommendations to be highly redundant.

In conclusion, the analysis of the re-ranking strategies shows that:

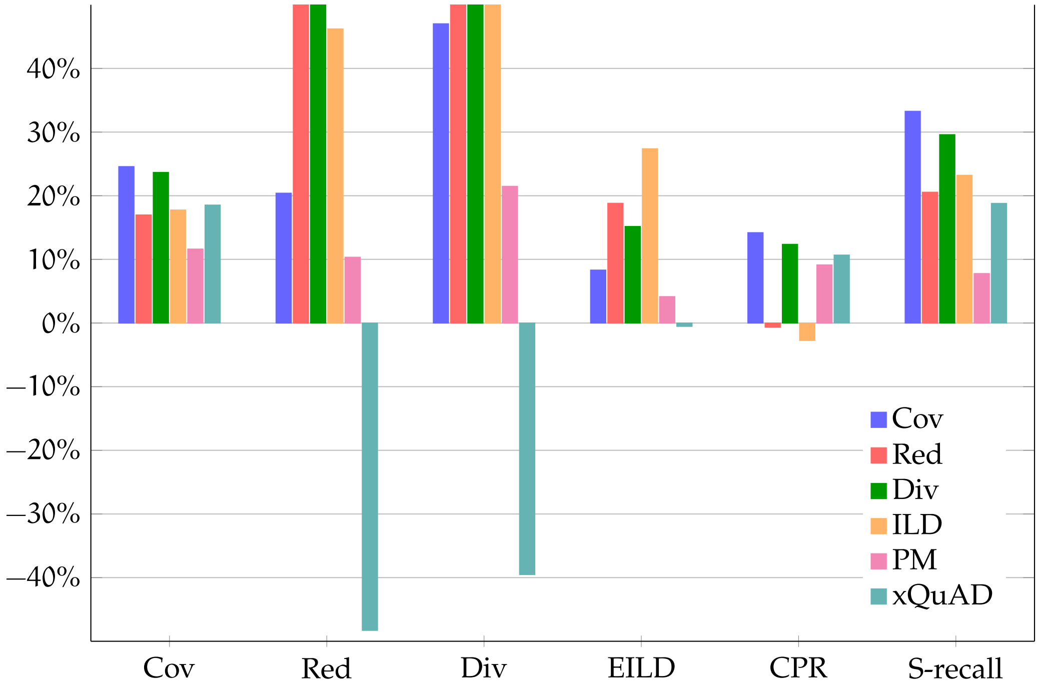

Figure 6.4: Relative difference over the iMF baseline recommender of the Binomial, PM, MMR and xQuAD diversifiers in MovieLens1M.

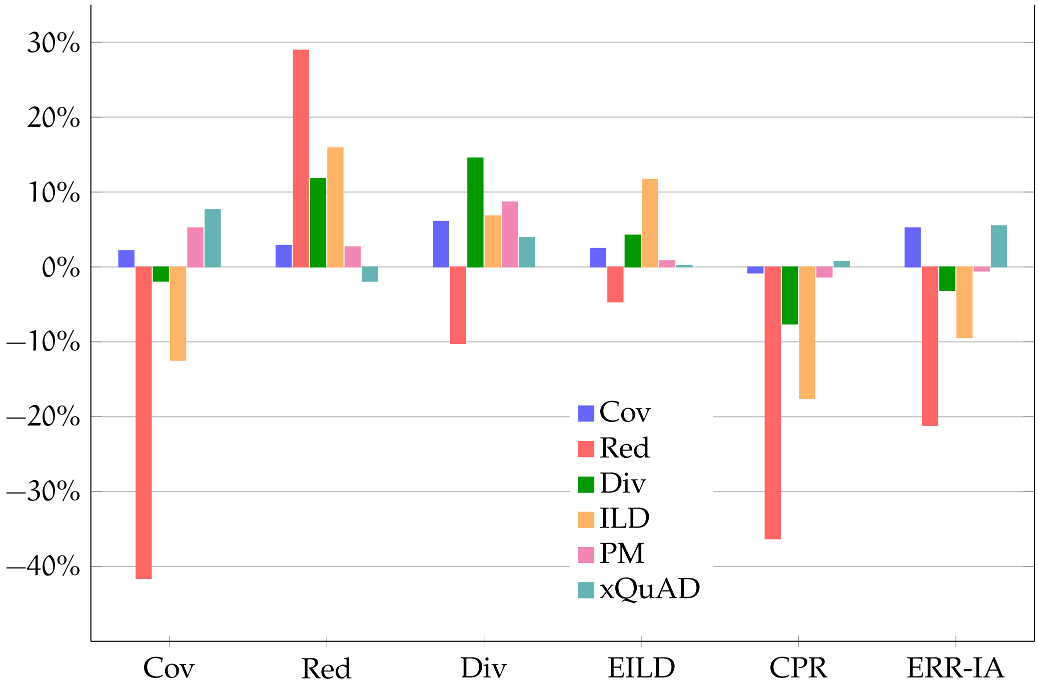

Figure 6.5: Relative difference over the iMF baseline recommender of the Binomial, PM, MMR and xQuAD diversifiers in MovieLens1M.

	Cov			**Red**			**Div**		
	0.0	**0.5**	**1.0**	**0.0**	**0.5**	**1.0**	**0.0**	**0.5**	**1.0**
Rnd	0.7259	0.7184	0.6750	0.6597	0.6140	0.5467	0.4816	0.4448	0.3748
Pop	0.8477	0.8392	0.8090	0.2767	0.2321	0.1800	0.2362	0.1973	0.1489
iMF	0.7351	0.7858	0.8170	0.2712	0.3582	0.4225	0.2104	0.2875	0.3477

Table 6.13: Variation of the α parameter that regulates the trade-off between genre generality and user relative importance for the iMF baseline in MovieLens1M.

- The Pair-Wise framework relates to our Binomial redundancy.
- The Proportionality Framework optimizes mainly genre coverage, although our adaptation to recommendation does not get satisfactory results when considering the relevance of recommendations.
- Intent-Aware metrics and diversification techniques favor retrieving items covering many genres regardless of their redundancy.

6.6.3 *Analysis of the Properties of the Binomial Framework*

In the last part of our experiments, we want to illustrate the unique properties of our Binomial framework, mainly the trade-off between generality and user relative relevance of genres and the size-awareness of our metrics and re-ranking strategies.

In Table 6.13 we show the effect of varying the α parameter (set to 0.5 in previous experiments) to consider only the generality of the genres ($\alpha = 0.0$) or to take exclusively the user preferences for the genres into account ($\alpha = 1.0$). We show the variation of this metric for the two non-personalized recommendations (random and most-popular) and iMF as representative of the personalized recommendations for the MovieLens1M dataset. The results clearly illustrate how, depending on the degree of personalization of the genre probability, the results favor the algorithm that takes the user preferences into account. As we can observe, when $\alpha = 0.0$ iMF is the worst option (or close to it) in the coverage and redundancy scores, and thus in the Binomial Diversity metric. However, the more we account for the user tastes into account – which are expected to be captured by the personalized nature of iMF – the better the personalized recommendation behaves with respect to the its non-personalized counterparts. These results provide an argument to consider, in any case, that genres should not be treated carelessly, since the preferences of each user clearly affect the way they perceive coverage and redundancy.

Finally, in order to evaluate the size-awareness of our Binomial framework, Table 6.14 shows the correspondence, using the iMF baseline in MovieLens1M, be-

	norel			**rel**		
	5	**10**	**20**	**5**	**10**	**20**
iMF	0.3895	0.3213	0.2875	0.3850	0.2875	0.1906
5	0.8926	0.5512	0.3776	0.4395	0.3029	0.1952
10	0.8466	0.8702	0.5183	0.4357	0.3348	0.2015
20	0.7766	0.8322	0.8557	0.4252	0.3285	0.2184

Table 6.14: Results at different cut-offs ($N = 5, 10, 20$) for the Binomial Diversity metric (without and with relevance) and its re-ranking in MovieLens1M.

tween the cut-off of the binomial diversification algorithm (the N in Equation 6.9) and the cut-off of the binomial diversity metric. For each diversification cut-off, the results correspond to the best λ of the objective function. As expected, the best diversification cut-off always agrees with the cut-off of the diversity metric, in both relevance-unaware and aware variants. This shows that our approach is able to leverage knowledge of the desired result set size in order to bring an additional made-to-fit improvement at the targeted cut-off, a feature that is not supported in any prior framework.

6.7 CONCLUSIONS

We tackle in this chapter the problem of diversity using genre information in Recommender Systems. An analysis of the properties of genres helps us define the requirements that a genre-based definition of diversity in recommendation should satisfy, namely coverage, redundancy and recommendation list size-awareness. We propose a Binomial framework that satisfies these properties. A metric is defined upon this framework, and a greedy re-ranking algorithm that optimizes it. Experiments with movie and music recommendation datasets validate the consistency of our framework, illustrate its properties, and show they comply with the stated requirements.

7

IMPROVING SALES DIVERSITY BY RECOMMENDING USERS TO ITEMS

7.1 INTRODUCTION

Sales Diversity (see Section 2.3.3.5) has been pointed out as a relevant quality of recommendation from the business point of view (Fleder and Hosanagar, 2009). It consists in "making the most of the catalog", that is, procuring that all or most products in the business catalog get purchased to some extent, rather than having sales concentrating around a few items. Sales Diversity gets meaning in the context of recommendation in the sense that recommending a product exposes it to being sold. By linking recommendation to purchase in the analysis of diversity, Sales Diversity becomes a shorthand for "promoting sales diversity".

As seen in Section 2.3.3.5, Sales Diversity can be measured by different metrics, such as Aggregate Diversity (Adomavicius and Kwon, 2012), Inter-User Diversity (Bellogín et al., 2010; Zhou et al., 2010), Entropy (Patil and Taillie, 1982), the Gini Index (Fleder and Hosanagar, 2009) and the rank and relevance-aware generalization of these metrics in our unified framework of Chapter 4.

Prior research (Adomavicius and Kwon, 2012) – confirmed by our experiments in Section 4.9 – has found there is an indirect connection between Sales Diversity and Long Tail Novelty (see Section 2.3.3.1): promoting long tail novel items has a positive effect on Sales Diversity, even though Sales Diversity and Long Tail Novelty are not in themselves the same thing. This effect is explained by the *popularity bias* in recommendations (see Section 2.2.4): collaborative filtering algorithms rely on previous interactions between users and items (in the form of ratings, play counts, and other kinds of feedback) to generate recommendations. In such interactions it is common to find a long tail effect, in which a reduced set of the most-popular items – the *short head* – account for a elevated proportion of the interactions of the users, while the much bigger *long tail* consists of less known items. Such long tail effect naturally influences the recommendation algorithms towards recommending the items in the short head, and therefore resulting in low Sales Diversity. The research and findings we report here provide means to better cope, directly and indirectly, with this bias, as we discuss in the next sections.

In this chapter, we consider a different outlook on the problem of Sales Diversity. If we aim to procure a fair opportunity for most items to be recommended, one may consider selecting which users each item should be recommended to, instead of the other way around. This view entails a symmetric swap of the recommendation task, whereby users are recommended to items rather than the opposite. From this perspective, we address three main questions:

- How can we define suitable and effective algorithms that recommend users to items?
- Does the inverted formulation actually enable any improvements in Sales Diversity?
- If so, what trade-offs if any are involved with respect to other qualities such as precision or recommendation novelty from the user point of view?

To address these questions we propose, firstly, to explore the application of state of the art collaborative filtering algorithms to the inverted recommendation task, that is, simply swapping the role of items and users in the algorithms. We find interesting derivations, equivalences, and new insights on the behavior of neighborhood-based algorithms in particular, where the inversion results in the emergence of new neighbor selection policies, with an impact on the potential connections to item popularity. We furthermore find that the inversion approach results in a significant increase of Sales Diversity while retaining a good trade-off on top-N item recommendation precision. In addition to this, we develop a probabilistic scheme which formulates user recommendation to items as a Bayesian layer which can be applied on top of any recommendation algorithm. The probabilistic scheme provides a principled means to isolate the item popularity component of the baseline algorithm; by means of simple smoothing techniques, the presence of this popularity component can be calibrated (i.e. kept unchanged or neutralized) to any desired degree. This parametrization is shown to enable an enhanced precision-diversity trade-off, even above, somewhat surprisingly, direct optimization approaches targeting the precision vs. Long Tail Novelty trade-off. Furthermore, the resulting algorithmic scheme is competitive with respect to direct optimization even in terms of Long Tail Novelty.

The rest of the chapter is structured as follows. Section 7.2 introduces the inverted recommendation task of recommending users to items. Then, Section 7.3 describes our first proposal of using inverted neighborhoods to improve Sales Diversity. An analysis of the properties of standard and inverted neighborhoods and their differences is shown in Section 7.4. Section 7.5 presents our second proposal, a probabilistic reformulation layer that allows the calibration of the popularity bias in state-of-the-art collaborative filtering algorithms. Experiments with two different recommendation scenarios – movies and music – are described in Section 7.6. Finally, Section 7.7 offers the conclusions.

The contents of this chapter have been presented in following published work:

- Vargas, S., Baltrunas, L., Karatzoglou, A., and Castells, P. (2014). Coverage, redundancy and size-awareness in genre diversity for recommender systems. In *Proceedings of the 8th ACM Conference on Recommender Systems*, RecSys '14, pages 209–216, New York, NY, USA. ACM
- Vargas, S. and Castells, P. (2014b). Vecindarios inversos para la mejora de la novedad en filtrado colaborativo. In *III Congreso Español de Recuperación de Información*, CERI'14

7.2 RECOMMENDING USERS TO ITEMS: PROBLEM STATEMENT

In this section we present the problem of recommending users to items. To abbreviate, throughout this chapter we define this problem as *inverted recommendation*, as opposed to the *standard recommendation* one. This problem swaps the roles of users and items and focuses on finding the users that would be more interested in a given item. Therefore, we prioritize the task of assigning each item in the catalog to appropriate users, thus diminishing and controlling the contribution of the popularity bias in the recommendation process. The idea of finding appropriate users for an item is partly inspired in the online dating recommendation problem (Pizzato et al., 2010) – consisting in finding adequate matches between users – in which it is equally important to provide recommendations relevant to the user as that the recommended contacts find the user relevant.

As mentioned in the introduction, we do not consider this task as an end by itself, but a means to improve Sales Diversity in the standard recommendation. In fact, starting from the inverted recommendation scenario, we propose two different approaches for the standard recommendation, namely **inverted nearest neighbors** and a **probabilistic reformulation layer**, that are capable of offering accurate recommendations with substantial improvements in Sales Diversity.

Analogously to the standard recommendation task, the inverted recommendation task can be formulated as defining a scoring function $\tilde{s} : \mathcal{I} \times \mathcal{U} \to \mathbb{R}$ which induces a ranking of users by their decreasing predicted relevance to item i. In a pure collaborative filtering setting, the input data for this task consist of the transposed rating matrix $\tilde{\mathcal{R}} = \mathcal{R}^t$.

Since collaborative filtering algorithms do not depend on the content or internal characteristics of both users and items, they can be adapted for this task without any modification apart from the change of roles between users and items. An initial observation is that, for many popular, state-of-the-art collaborative filtering algorithms, the scoring function $\tilde{s}$ is actually identical to that of the original problem s. That is the case, for instance, of many matrix factorization approaches, such as the implicit matrix factorization of Hu et al. (2008):

$$\tilde{s}_{iMF}(i,u) = Q_i\, P_u^t = P_u\, Q_i^t = s_{iMF}(u,i) \tag{7.1}$$

There are other collaborative filtering approaches that break this symmetry. That is the case of other matrix factorization approaches (Shi et al., 2012a, 2013; Takács and Tikk, 2012) which, even having the same scoring function as in Equation 7.1, have non-interchangeable roles for users and items in their model training, and thus provide new scoring functions between users and items. However, we focus on the case of nearest neighbors approaches, whose asymmetry offers an interesting, new alternative for generating diverse recommendations.

7.3 INVERTED NEAREST NEIGHBORS

A first application of the inverted recommendation task lies in the asymmetry of the nearest neighbor approaches when applied to the inverted recommendation task. Throughout this section we focus on the user-based nearest neighbors approach, since most of the observations, unless explicitly discussed, are straightforwardly translatable to the item-based alternative.

The scoring functions of the user-based and item-based nearest neighbors recommenders (Aiolli, 2013; Cremonesi et al., 2010) can be formulated as follows:

$$s_{UB}(u,i) = \sum_{v \in \mathcal{U}} \mathbb{1}_{v \in N(u)}\, \mathrm{sim}(u,v)\, r_{v,i} \tag{7.2}$$

$$s_{IB}(u,i) = \sum_{j \in \mathcal{I}_u} \mathbb{1}_{i \in N(j)}\, \mathrm{sim}(i,j)\, r_{u,j} \tag{7.3}$$

where $\mathrm{sim}(u,v)$ is a similarity function between two users and $N(u)$ is the neighborhood of user u, containing the topmost similar users to item u.

Reformulating these algorithms in the inverted recommendation task, the scoring functions become:

$$\tilde{s}_{UB}(i,u) = \sum_{j \in \mathcal{I}} \mathbb{1}_{j \in N(i)}\, \mathrm{sim}(i,j)\, \tilde{r}_{j,u} \tag{7.4}$$

$$\tilde{s}_{IB}(i,u) = \sum_{v \in \mathcal{U}_i} \mathbb{1}_{u \in N(v)}\, \mathrm{sim}(u,v)\, \tilde{r}_{i,v} \tag{7.5}$$

As previously commented, the symmetry of the nearest neighbors scoring functions with respect to the original problem is broken. In particular, the user-based approach s_{UB} in Equation 7.2 is notably different from the scoring function $\tilde{s}_{UB}$ in Equation 7.4. Interestingly, it is almost equivalent to the item-based approach $\tilde{s}_{IB}$ for the inverted recommendation in Equation 7.5, the difference lying on the neighborhood selection criterion, i.e., $\mathbb{1}_{u \in N(v)}$ against $\mathbb{1}_{v \in N(u)}$. Actually, one can re-formulate the scoring function $\tilde{s}_{IB}$ as a variant of the standard user-based ap-

sim	u_1	u_2	u_3	u_4
u_1	-	0.1	0.3	0.4
u_2	0.1	-	0.5	0.0
u_3	0.3	0.5	-	0.6
u_4	0.4	0.0	0.6	-

u	$N_2(u)$	$N_2^{-1}(u)$
u_1	$\{u_3, u_4\}$	$\{u_2, u_4\}$
u_2	$\{u_1, u_3\}$	$\{u_3\}$
u_3	$\{u_2, u_4\}$	$\{u_1, u_2, u_4\}$
u_4	$\{u_1, u_3\}$	$\{u_1, u_3\}$

Table 7.1: Example of user neighborhoods of size 2.

proach s_{UB} in which the policy for neighbor selection is inverted, that is, by considering user *inverted neighborhoods* $N^{-1}(u)$ defined as

$$N^{-1}(u) = \{v \in \mathcal{U} \ : \ u \in N(v)\} \tag{7.6}$$

where $N(v)$ is the original neighborhood for a user v, so that $\mathbf{1}_{u \in N(v)} = \mathbf{1}_{v \in N^{-1}(u)}$. The concept of inverted neighborhoods originally appeared in (Sarwar et al., 2001), where it was proposed as an ad-hoc method to efficiently predict ratings for item-based approaches, without any relationship with the inverted recommendation tasks or the improvement of Sales Diversity.

Note what the resulting inverted neighborhood formation policy means: instead of selecting $N(u)$ as the top-K most similar users to the target user u, all the users v for which the target user is among the K most similar to v are selected as the neighbors $N^{-1}(u)$ of u. Table 7.1 shows an example of a community of users and their corresponding standard and inverted neighborhoods for $K = 2$. This has several consequences. In the first place, the resulting, inverted neighborhoods no longer have all the same size. The size of the inverted neighborhood of a user u is the number of users to whose neighborhood u belongs – in particular this means that some users might have an empty neighborhood at the cost of user coverage of the recommendation, but we have observed in our experiments that this situation does not happen in practice if the original neighborhoods are large enough. Having different neighborhood sizes is not necessarily a drawback, on the contrary, it may be favorable that users have as large a neighborhood as the reliability of the available data for each user enables.

The inverted neighborhoods approach implies, on the other hand, that all users will appear in exactly the same number $|N(u)|$ of inverted neighborhoods (except perhaps a few low activity users for which it was not possible to form a direct neighborhood of size K in the first place). This flattens the influence power of all users, so that all users' opinions "count" to the same extent overall in the produced recommendations. This may be expected to avoid a concentration of recommendations over the tastes of a reduced set of users, thereby indirectly enhancing a more even distribution of items across recommendations to the user population.

	$\text{avg}_{u \in \mathcal{U}} \lvert \mathcal{I}_u \rvert$	$\text{avg}_{i \in \mathcal{I}} \lvert \mathcal{U}_i \rvert$
ML1M	165.60	269.89
Netflix	209.25	5,654.50
MSD	43.03	127.88

Table 7.2: Average user and item profile size in MovieLens1M, Netflix and Million Song Dataset.

In the case of the item-based variant this effect is even more direct: if all items appear in the inverted neighborhood of the same number of items (neighbor items being the candidates for recommendation in the item-based nearest neighbors method), they will have more even chances of making it to the top-N of recommendations, whereby one may expect better distributed recommendations over the set of items (i.e. more diverse "sales"). Moreover, long tail items, by getting a more equal opportunity to be recommended with respect to popular items, might make for a Long Tail Novelty enhancement of recommendations.

In order to have a preliminary understanding of these potential effects, we analyze more closely in the next section the relation between user and item characteristics, namely profile size, and their distribution across neighborhoods for the direct and inverted selection policies. Our discussion of the potential effects on final recommendation diversity is so far speculative and needs to be tested empirically, as we report in Section 7.6.1.

7.4 NEIGHBORHOOD BIAS ANALYSIS

We test and illustrate the biases suggested in the previous section by taking some measurements on data from the MovieLens1M, Netflix Prize and Million Song datasets, for which we study the characteristics of user and item neighborhoods with different neighborhood sizes K and using cosine similarity as in (Cremonesi et al., 2010; Aiolli, 2013).

We show in Tables 7.4 and 7.5 the following measurements:

- Average profile size of the neighbors (**S**), in order to detect any possible bias towards neighbors with profile sizes significantly different from the average profile size (displayed in Table 7.2).

- Gini coefficient of the user or item distribution across neighborhoods (**G**), as an indicator to detect any imbalance in the distribution of the number $|N_K^{-1}|$ of neighborhoods a user or item belongs to.

- Correlation between the profile size of a user or item and the number of neighborhoods she or it belongs to (**C**), to see if an existing imbalance is

	user-based	item-based
S	$\underset{u}{\text{avg}} \underset{v \in N(u)}{\text{avg}} \lvert \mathcal{I}_v \rvert$	$\underset{i}{\text{avg}} \underset{j \in N(i)}{\text{avg}} \lvert \mathcal{U}_j \rvert$
G	$1 - \mathrm{Gini}(\lvert N^{-1}(u) \rvert)$	$1 - \mathrm{Gini}(\lvert N^{-1}(i) \rvert)$
C	$\rho(\lvert \mathcal{I}_u \rvert, \lvert N^{-1}(u) \rvert)$	$\rho(\lvert \mathcal{U}_i \rvert, \lvert N^{-1}(i) \rvert)$

Table 7.3: Definition of the neighborhood properties. N denotes in this case a generic user or item neighborhood, either standard or inverted.

caused for a bias in the profile size of users or items in the neighbor selection process.

A more formal definition of the above measurements is given in Table 7.3, where we denote by N_K the direct neighborhood formed by the K most similar users or items, and by N_K^{-1} the inverted neighborhood.

The results in Table 7.4 reveal, as hypothesized, biases and concentrations in the selection of user for standard neighborhoods. In the case of MovieLens and Netflix data, the standard neighborhood method is slightly biased towards selecting users with large profiles and shows a clear concentration on a small subset of users. In the case of the Million Song Dataset, there is an opposite bias towards small profiles, which also causes a concentration of neighbors. A possible explanation of why these methods differ in the direction of the bias may lie in the incomparable number of items between them and the different levels of sparsity in each dataset. In any case, the inverted selections strategy corrects this biases, that is, eliminates the correlation between profile size and the number of neighborhoods a user belongs to and, at the same time, creates a perfectly balanced distribution of this number.

Table 7.5 shows the equivalent measurements for item neighborhoods. Again, we can observe biases and concentration in the direct selection method that are partly solved by the inverted neighborhoods. The MovieLens data shows a bias towards popular items that, ultimately, compose the majority of the neighborhoods. The inverted item neighborhoods, as in the user neighborhood case, alleviate this problem by allowing more, less popular items appear in the neighborhoods. The Netflix data also shows similar biases that are solved by the inverted item neighborhoods, which achieve a perfectly equitative distribution of the items in the neighborhoods, doing away with the bias towards popular items. In the Million Song dataset, a bias towards popular items is also observed for large neighborhood sizes, and an uneven distribution of the items in the distribution is observed for all neighborhood sizes. Again, inverted neighborhoods help solving these effects by significantly reducing the bias towards popular items and achieving more uniform distributions in the number of neighborhoods each item belongs to.

		S		**G**		**C**	
	K	N_K	N_K^{-1}	N_K	N_K^{-1}	N_K	N_K^{-1}
ML1M	10	202.70	90.56	0.41	1.00	0.34	0.03
	20	209.20	87.12	0.45	1.00	0.40	0.04
	50	219.20	84.42	0.49	1.00	0.50	0.07
	100	227.79	83.45	0.51	1.00	0.59	0.09
	200	236.43	83.87	0.52	1.00	0.67	0.12
	500	244.46	87.41	0.52	1.00	0.75	0.17
	1,000	240.87	94.47	0.53	1.00	0.77	0.22
	2,000	217.85	108.12	0.59	1.00	0.75	0.03
	5,000	152.03	132.48	0.87	0.99	0.52	0.20
Netflix	10	286.39	137.91	0.18	1.00	0.07	-
	20	291.07	131.91	0.20	1.00	0.09	-
	50	298.20	124.68	0.22	1.00	0.11	0.00
	100	304.38	120.08	0.25	1.00	0.12	0.00
	200	311.24	116.46	0.27	1.00	0.15	0.00
	500	321.53	113.25	0.30	1.00	0.18	0.01
	1000	330.29	111.58	0.33	1.00	0.22	0.01
	2000	339.96	110.41	0.35	1.00	0.26	0.02
	5000	353.90	110.12	0.21	1.00	0.32	0.02
MSD	10	23.12	35.41	0.18	1.00	-0.11	0.00
	20	24.29	35.58	0.21	1.00	-0.11	0.00
	50	26.27	35.96	0.26	1.00	-0.12	0.00
	100	28.07	36.41	0.30	1.00	-0.12	0.00
	200	30.06	37.07	0.34	1.00	-0.12	0.01
	500	32.75	38.63	0.40	1.00	-0.11	0.01

Table 7.4: Properties of user neighborhoods with cosine similarity for the MovieLens1M, Netflix and Million Song datasets. Dashes mark undefined correlations since $|N^{-1}(u)|$ was constant. See Table 7.3 for the meaning of S, G and C.

		S		G		C	
	K	N_K	N_K^{-1}	N_K	N_K^{-1}	N_K	N_K^{-1}
ML1M	10	347.22	142.80	0.47	1.00	0.32	-
	20	366.25	133.50	0.51	1.00	0.44	0.02
	50	392.59	127.67	0.55	1.00	0.61	0.05
	100	413.01	124.64	0.56	0.99	0.72	0.07
	200	433.26	123.46	0.57	0.99	0.79	0.08
	500	446.91	130.16	0.55	0.98	0.84	0.11
	1,000	423.92	150.92	0.56	0.97	0.83	0.16
	2,000	350.15	193.13	0.67	0.94	0.70	0.22
	3,706	275.64	275.64	0.87	0.87	0.46	0.46
Netflix	10	7,663.08	2,885.83	0.48	1.00	0.13	-
	20	8,172.21	2,665.99	0.48	1.00	0.17	-
	50	9,069.98	2,410.01	0.48	1.00	0.25	-
	100	9,848.19	2,257.51	0.48	1.00	0.31	-
	200	10,629.88	2,137.27	0.48	1.00	0.38	-
	500	11,335.30	2,076.77	0.49	1.00	0.45	-
	1,000	11,229.95	2,229.19	0.50	1.00	0.47	-
	2,000	9,975.00	2,695.19	0.54	1.00	0.46	0.00
	5,000	6,812.03	3,929.65	0.69	1.00	0.31	0.00
MSD	10	143.90	116.64	0.60	1.00	0.02	0.01
	20	172.99	110.90	0.64	1.00	0.05	0.01
	50	240.60	105.88	0.67	0.99	0.12	0.02
	100	332.47	103.86	0.70	0.97	0.22	0.04
	200	480.62	113.86	0.70	0.94	0.42	0.05
	500	794.93	154.76	0.62	0.86	0.64	0.08
	1,000	1,109.70	211.77	0.53	0.78	0.71	0.11
	2,000	1,425.57	295.69	0.45	0.67	0.75	0.15
	5,000	1,732.61	451.97	0.37	0.52	0.76	0.22

Table 7.5: Properties of item neighborhoods with cosine similarity for the MovieLens1M, Netflix and Million Song datasets. Dashes mark undefined correlations since $|N^{-1}(i)|$ was constant. See Table 7.3 for the meaning of S, G and C.

7.5 PROBABILISTIC REFORMULATION LAYER

The inverted recommendation task can also be addressed in probabilistic terms. Probabilistic formulations have been used extensively in the conventional item recommendation task as a means to develop collaborative filtering methods. For instance, Hofmann (2004) proposed ranking items by the decreasing probability $p(i|u)$ that the target user would prefer each item over the others. This principle is developed by means of an adaptation of the probabilistic Latent Semantic Analysis (pLSA) framework into an effective scoring procedure for producing ranked recommendations.

Turning the task around, recommending users for items would consist in estimating $p(u|i)$ for each user u given an item i. A straightforward way of linking any recommendation algorithm to a probabilistic formulation can be established by assuming that the recommender scoring function $s(u,i)$ is roughly proportional to $p(u,i)$. This assumption, coarse as it may be, provides a very direct means to bring the recommendation algorithm to a probabilistic interpretation as per:

$$p(u\,|\,i;s) \sim \frac{s(u,i)}{\sum_v s(v,i)} \tag{7.7}$$

We can therefore employ this approach to obtain an inverted recommendation method out of any direct item recommendation algorithm. Note that the resulting formulation produces a totally equivalent output as its scoring function preserves the original ranking, i.e., $p(u\,|\,i;s) \propto s(u,i)$. The formulation is however useful as it enables a probability-based manipulation of the popularity bias in recommendation algorithms, as we see next.

First, the resulting output of the inverted recommendation should be reverted to a list of ranked items to be delivered to each user. A principled way to do this is by applying Bayesian inversion on $p(u|i)$, thereby obtaining an estimate for $p(i|u)$ as a suitable scoring function for ranking items for each user:

$$p(i\,|\,u;s) = \frac{p(u\,|\,i;s)\,p(i;s)}{\sum_j p(u\,|\,j;s)\,p(j;s)} \tag{7.8}$$

where the prior $p(i;s)$ represents how likely the item is to be the favorite of a random user.

Note that we could instead have derived an estimate of $p(i\,|\,u;s)$ by an equivalent symmetric version of Equation 7.7. However, the advantage of Equation 7.8 is that it explicitly reflects the popularity component carried by the item prior $p(i;s)$. Using the same assumption as before between the scoring function and probabilities, we have:

	UB	IB	iMF	pLSA
ML1M	0.99	0.80	0.98	0.99
Netflix	0.99	0.74	0.94	0.99
MSD	0.89	0.80		

Table 7.6: Pearson correlation between item popularity and the item priors $p(i; s, 0)$ of the baseline recommendation algorithms in MovieLens1M, Netflix and Million Song Dataset.

$$p(i; s) \sim \frac{\sum_u s(u,i)}{\sum_j \sum_u s(u,j)} \tag{7.9}$$

Now that the popularity component is isolated, we propose to smooth the prior estimate in a way that has it range from the literal estimate based on the recommender's scores, to a flat uniform background prior where all items are considered equally popular. We do so by an entropic regularization of the estimate – similar to the tempered expectation maximization algorithm in (Hofmann, 2004), which simply introduces an exponent in the expression:

$$p(i; s, \alpha) \sim \frac{\left(\sum_u s(u,i)\right)^{1-\alpha}}{\sum_j \left(\sum_u s(u,j)\right)^{1-\alpha}} \tag{7.10}$$

In the above expression, the $\alpha \in [0, 1]$ smoothing parameter allows controlling how much of the algorithm popularity bias we wish to leave as is or remove.

Interestingly, by combining equations 7.8 and 7.10, the resulting probabilistic interpretation $p(i \,|\, u; s)$ can be simplified to a re-scoring procedure for a standard scoring function s as follows:

$$s_{PR}(u,i) = s(u,i) \left(\sum_v s(v,i) \right)^{-\alpha} \tag{7.11}$$

This last reformulation allows to see more clearly the role of the parameter α. On one hand, when $\alpha = 0$ we obtain the original recommendation list created with the scoring function. On the other hand, when $\alpha = 1$ the prior is uniform and thus the recommendations to users will be completely based on $p(u \,|\, i; \tilde{s})$, eliminating any possible popularity bias in the items. The use of intermediate values of α is a means to provide more varied recommendation lists while retaining an appropriate level of relevance, that is α is a parameter that controls the relevance/novelty tradeoff, only that novelty is not applied as the opposite to popularity (as is the case in most novelty enhancement approaches (Adomavicius and Kwon, 2012)), but rather as neutrality with respect to popularity.

To illustrate how we can control the popularity bias by the proposed approach, we show in Table 7.6 the Pearson correlation values between the priors $p(i; s, 0)$ and the popularity of the items (understood as the number of users who know – i.e. have rated – the item) for the recommendations baselines of our experimental design (see Chapter 3) for the MovieLens1M, Netflix and Million Song datasets. The values reveal a strong correlation between our score-based estimate of the item priors and the actual popularity of the items. This, on the other hand, empirically illustrates the popularity bias of these state of the art algorithms as discussed in Section 2.2.4, and shows how the prior component captures it, enabling its gradual adjustment by the α parameter.

7.6 EXPERIMENTS

In order to test the effectiveness and analyze the properties of the proposed inverted nearest neighbor methods and the probabilistic reformulation layer, we carry out experiments in the context of our experimental design detailed in Chapter 3. These experiments aim to answer the following questions:

- Do the properties of the inverted neighborhoods observed in Section 7.4 translate to improvements in terms of Sales Diversity and Long Tail Novelty compared with standard neighborhoods?
- How does the probabilistic reformulation layer compare to direct long tail optimization methods?

7.6.1 *Evaluation of the Inverted Nearest Neighbors*

For answering the first question, we first compare the inverted nearest neighbors approaches described (Equations 7.4 and 7.5) to the corresponding standard user-based and item-based formulations (Equations 7.2 and 7.3) for different neighborhood sizes K. Results are evaluated in precision (P), the rank and relevance-unaware version of the Expected Popularity Complement (EPC in Equation 4.26) and the complement of the Gini Index (1−Gini) at cut-off 20. Second, in order to provide a wider perspective in the context of alternative recommendation algorithms and metrics, we compare both inverted and standard neighborhood methods with the rest of recommendation baselines of our experimental design using the previous metrics as well as the Gini-Simpson Index (GSI), Aggregate Diversity (Aggr-div) and Entropy (H), which are (rank and relevance-unaware) equivalent to the Sales Diversity metrics of our framework of Chapter 4.

		P	EPC	1−Gini	GSI	Aggr-div	H
ML1M	**Rnd**	0.0057	0.9668	0.8873	0.9997	3,680.2	11.8197
	Pop	0.1215	0.6818	0.0089	0.9726	110.4	5.4309
	iMF	0.2335	0.8216	0.1027	0.9970	1,270.4	8.9841
	pLSA	0.2111	0.8131	0.0916	0.9965	1,122.4	8.8099
	UB (N_{100})	0.2056	0.7605	0.0430	0.9920	749.2	7.6988
	UB (N^{-1}_{100})	0.2173	0.8019	0.0904	0.9958	1,709.8	8.7310
	IB (N_{10})	0.1874	0.7738	0.0419	0.9912	1,014.0	7.6272
	IB (N^{-1}_{10})	0.1873	0.8166	0.1033	0.9965	1,788.6	8.9394
Netflix	**Rnd**	0.0022	0.9866	0.4769	0.9999	9,320.0	13.1673
	Pop	0.0909	0.7171	0.0018	0.9712	152.4	5.3711
	iMF	0.1842	0.8110	0.0164	0.9962	1,702.0	8.6149
	pLSA	0.1778	0.8433	0.0297	0.9975	3,912.4	9.4079
	UB (N_{100})	0.1923	0.7917	0.0124	0.9949	3,181.2	8.1939
	UB (N^{-1}_{100})	0.1627	0.8119	0.0539	0.9980	9,220.4	9.9996
	IB (N_{10})	0.1582	0.7852	0.0093	0.9909	6,023.2	7.5618
	IB (N^{-1}_{10})	0.1669	0.8244	0.0241	0.9971	7,753.0	9.0739
MSD	**Rnd**	0.0001	0.9998	0.3270	1.0000	150,818	17.1305
	Pop	0.0185	0.9582	0.0001	0.9516	36	4.4003
	UB (N_{200})	0.1018	0.9830	0.0157	0.9960	51,598	11.2881
	UB (N^{-1}_{200})	0.0909	0.9922	0.0371	0.9994	80,341	13.3122
	IB (N_{20})	0.1078	0.9925	0.0524	0.9992	104,769	13.7796
	IB (N^{-1}_{20})	0.1091	0.9933	0.0543	0.9994	105,953	13.9302

Table 7.7: Comparison of inverted neighborhood methods to other recommendation algorithms in MovieLens1M, Netflix and Million Song Dataset.

Figure 7.1 shows the comparison of the standard and inverted nearest-neighbor approaches in terms of the aforementioned metrics for different neighborhood sizes from 10 to 5000.

The results confirm a systematic increase in Sales Diversity (measured by the Gini coefficient), as hypothesized in Section 7.3. The improvement is consistent in the user-based and item-based versions for all neighborhood sizes on the three datasets. Notably moreover, for the item-based approach, the inverted method offers better accuracy and novelty for every value of K. For inverted user-based nearest neighbors, accuracy and novelty are better than in direct nearest neighbors

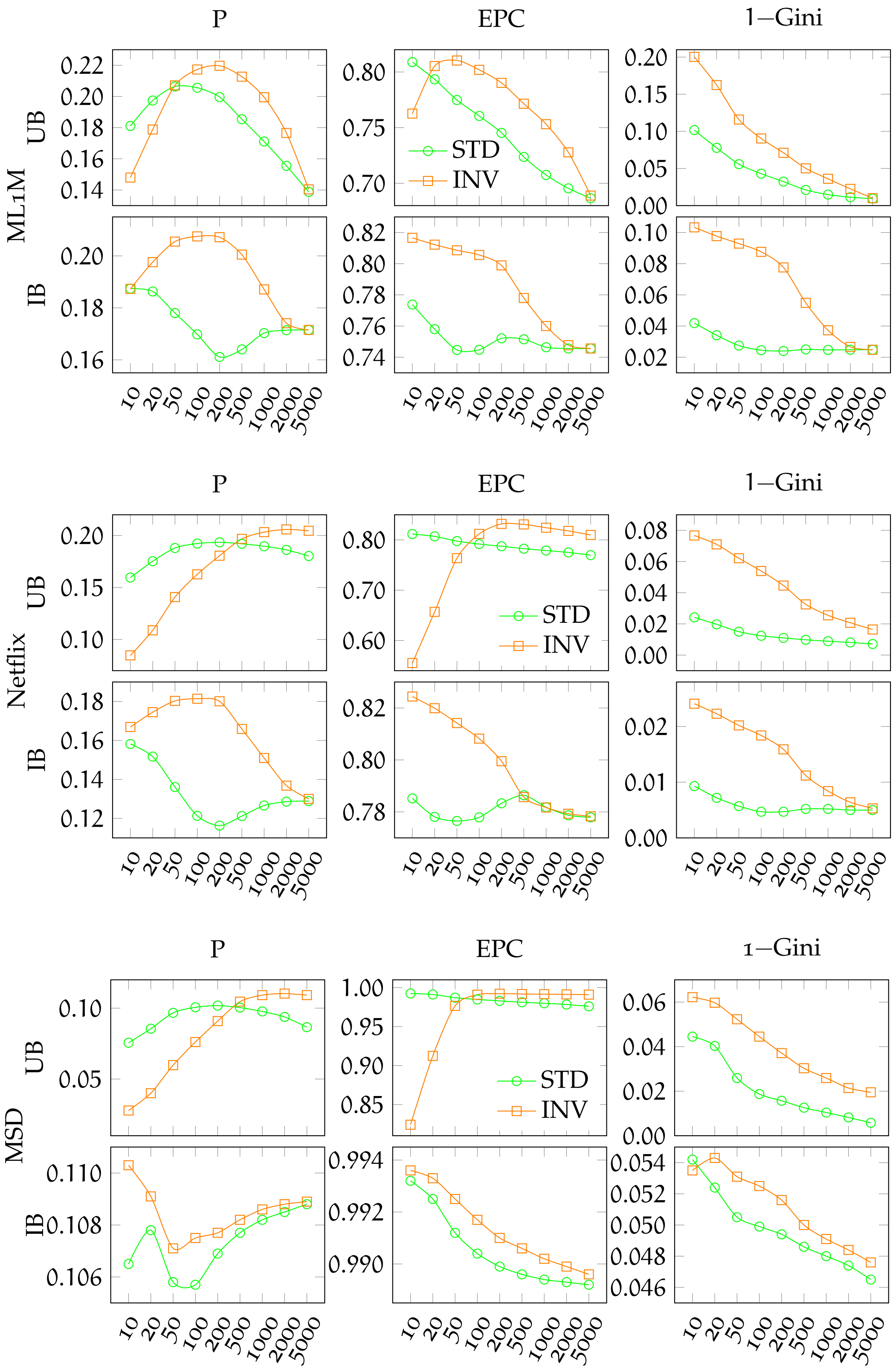

Figure 7.1: Results of the inverted nearest neighbors in MovieLens1M, Netflix and Million Song Dataset.

	K	**UB** (N_K)	**UB** (N_K^{-1})	**IB** (N_K)	**IB** (N_K^{-1})
ML1M	**10**	1.0000	0.8993	1.0000	1.0000
	20	1.0000	0.9656	1.0000	1.0000
	50	1.0000	0.9942	1.0000	1.0000
	100	1.0000	0.9991	1.0000	1.0000
	200	1.0000	0.9998	1.0000	1.0000
	500	1.0000	0.9999	1.0000	1.0000
	⩾1,000	1.0000	1.0000	1.0000	1.0000
Netflix	**10**	0.9913	0.6342	0.9990	0.9989
	20	0.9978	0.7556	0.9991	0.9991
	50	0.9992	0.8873	0.9992	0.9992
	100	0.9992	0.9522	0.9992	0.9992
	200	0.9992	0.9857	0.9992	0.9992
	500	0.9992	0.9983	0.9992	0.9992
	⩾1,000	0.9992	0.9992	0.9992	0.9992
MSD	**10**	1.0000	0.8278	0.9995	0.9998
	20	1.0000	0.9170	0.9998	1.0000
	50	1.0000	0.9826	1.0000	1.0000
	100	1.0000	0.9979	1.0000	1.0000
	200	1.0000	0.9999	1.0000	1.0000
	⩾500	1.0000	1.0000	1.0000	1.0000

Table 7.8: User coverage as the fraction of users in the test split which receive a recommendation in the MovieLens1M, Netflix and Million Song datasets.

only for large enough neighborhoods ($K \geqslant 100$). This is caused by user coverage degradation that occurs with smaller K, that is, many users do not receive recommendations since their inverted neighborhoods are empty, resulting in a penalization in metrics such as precision and EPC (we sum zero precision and EPC when a user cannot be delivered a recommendation). The details about the user coverage degradation are shown in Table 7.8. In a real recommender system this would not be acceptable, and a fallback solution, such as using the standard neighborhood, should be resorted to in those cases. However, in this analysis we are interested in the properties of the pure algorithm, and we therefore report the results for a plain version of the approach.

Table 7.7 shows the comparison of the inverted and standard neighborhood algorithms with random, popularity-based and latent factors recommendations. As seen in Chapter 4, random recommendation, as expected, produces inaccurate but highly novel and diverse results in both datasets. Popularity-based recommendation also yields predictable outcomes, producing moderately accurate results – depending on the sparsity of each dataset – and the lowest possible Long Tail Novelty and Sales Diversity – which should be so by definition. Latent factors algorithms (in MovieLens1M and Netflix only) have very strong results in terms of precision, novelty and diversity. Standard nearest neighborhood algorithms present competitive results in terms of accuracy compared to latent factor algorithms but have, in general, significant lower values in terms on Long Tail Novelty and Sales Diversity. This weakness is solved with the introduction of inverted neighbors, which – for equivalent neighborhood profile sizes – offer clear improvements in terms of novelty and diversity, making them comparable with latent factors approaches.

7.6.2 *Evaluation of the Probabilistic Reformulation Layer*

Regarding the second question formulated in the beginning of this section, we run a comparison between our probabilistic reformulation approach in Equation 7.11 and the re-ranking strategy of Section 4.8 the with Popularity Complement item novelty model (Equation 4.2) – equivalent to one of the proposed methods in (Adomavicius and Kwon, 2012):

$$s_{nov} = (1 - \lambda)\ s(u, i) + \lambda\ nov_{PC}(i)$$

The two compared approaches have a parameter that controls the trade-off between the original scoring function ($\alpha = 0.0$ and $\lambda = 0.0$) and a pure novelty component ($\alpha = 1.0$ and $\lambda = 1.0$). We explore these trade-offs by a grid search on the full interval by steps of 0.1. Results are evaluated with precision (P), the rank and relevance-unaware version of the Expected Popularity Complement (EPC) and the complement of the Gini Index ($1 - Gini$) at cut-off 20.

The results of our experiments with the probabilistic reformulation layer approach are shown in Figure 7.2. For each dataset-baseline pair we display two scatter plots showing the trade-offs between precision vs. Long Tail Novelty and Sales Diversity for the novelty-oriented re-scoring technique (NOV) and our probabilistic approach (PR). Curves for each approach start from $\alpha = \lambda = 0.0$ as the points with the lowest novelty and diversity and, as α and λ tend to 1.0, improve in terms of EPC and Gini while – generally – having lower precision values. Assuming that the interpolated lines are a good approximation to the continuous range of the trade-off parameters, we determine that a method is better that the other if its curve is generally above the other in each plot. Under this criterion, the results in Figure 7.2 show the validity of our probabilistic approach.

In the MovieLens1M and the Netflix data, we can see how the compared approaches show practically identical trade-offs in terms on EPC and, in terms of

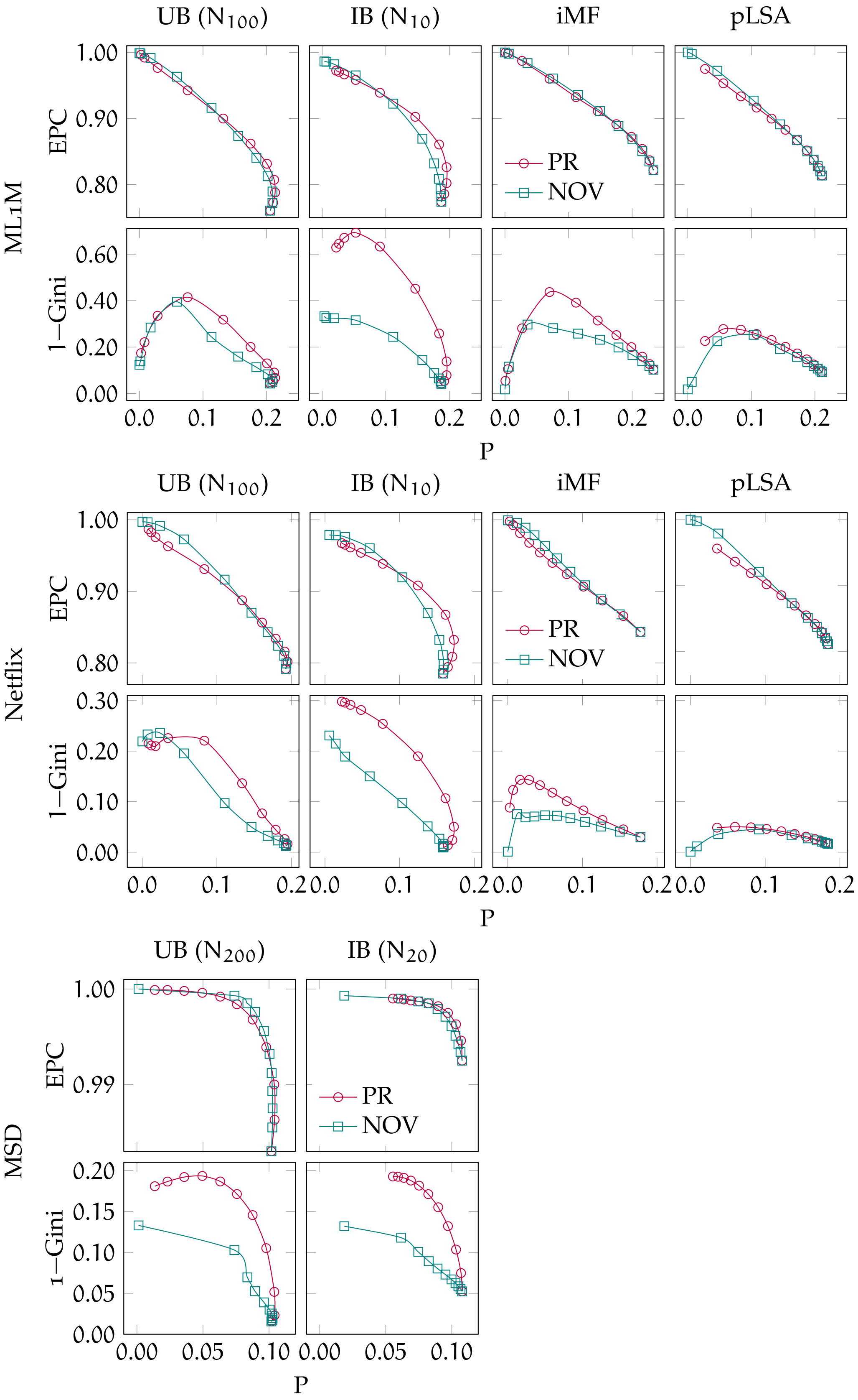

Figure 7.2: Results of the probabilistic reformulation layer in MovieLens1M, Netflix and Million Song Dataset.

Gini, our probabilistic method clearly outperforms the novelty-oriented re-ranking. Surprisingly for the item-based nearest neighbors, the probabilistic approach outperforms the original scoring function even in terms of precision and, among baselines, it is the one that achieves the highest Sales Diversity scores. On the other side, the improvements on the iMF matrix factorization approach, although being perceptible, are more limited than those in the nearest neighbors recommenders and, regarding pLSA, the proposed probabilistic reformulation offers practically identical results to the novelty-oriented reranking, notably an almost imperceptible improvement in terms of Sales Diversity.

In the Million Song dataset the results are analogous. Both re-ranking techniques present similar outcomes in terms of EPC, while the probabilistic approach clearly outperforms the novelty-oriented re-ranking. Again, item-based nearest neighbors is the baseline that enables a higher improvement in terms of Sales Diversity.

In conclusion, the proposed probabilistic approach provides a new method for improving trade-offs between accuracy, Long Tail Novelty and Sales Diversity. Compared to a simpler approach that optimizes directly the long-tail novelty of recommendations, our proposal obtains comparable results in terms of the novelty of recommendations, while it shows clearly better results in terms of Sales Diversity.

7.7 CONCLUSIONS

Starting from the aim of improving Sales Diversity by recommending users to items, we explore in this chapter where the inversion of the recommendation task leads to. By ranking users for items, the recommendation approach focuses on the relevance of user-item pairs in a different way, and item popularity gets left aside as a result. Starting from this task inversion, we derive two approaches to improve the Sales Diversity of the original item recommendation task. The first one, inverted neighborhoods, results in a novel way of "democratizing" the weight of user opinions (in the user-based approach) and item opportunity (in the item-based variant), leading to substantial improvements in terms of Sales Diversity, competitive results in Long Tail Novelty and a good precision trade-off. The second approach, a probabilistic reformulation of the recommendation problem, allows isolating the popularity component of any recommendation baseline and calibrate it in order to increase the chances of less popular items to appear in recommendations lists. The experiments conducted confirm and illustrate the effectiveness of our proposals.

The symmetric inversion of the recommendation task entails more than a simple transposition of the rating matrix. It brings up a new view on the task where the system seeks the most appropriate users to whom an item can be recommended, even though the final action is still the delivery of a ranked list of items to each user. This problem statement can reflect real-world situations where a business is selecting targets for advertising a particular product.

8

CONCLUSIONS AND FUTURE WORK

8.1 INTRODUCTION

We have developed in this thesis a series of principled approaches to evaluate and enhance the novelty and diversity of recommendations. First we have proposed a unified framework to explain and model several perspectives of novelty and diversity in recommendations. Then, an adaptation of the diversity metrics and diversification methods in Information Retrieval to Recommender Systems, specifically the so-called Intent-Aware framework, has been made. Beyond this, a recommendation-specific proposal to measure genre-based diversity in recommendations has been presented. Finally, we have addressed the popularity bias and his effect on Sales Diversity by developing two approaches based on a partial inversion of the recommendation task by selecting users for items instead of the opposite.

In this chapter we present the main conclusions derived from our contributions, and perspectives for future work. In Section 8.2 we present a summary of this work and its resulting contributions. In Section 8.3 we provide the future research directions for expanding our work.

8.2 SUMMARY AND CONTRIBUTIONS

We summarize and highlight next the main findings and contributions of this thesis, addressing the research goals stated in Chapter 1.

8.2.1 *A Unified Framework for Novelty and Diversity*

In Chapter 4 we have proposed a unified framework to assess novelty and diversity in Recommender Systems. By modeling the effective novelty and diversity of recommendation lists by means of an item novelty model and a browsing model, we provide a common scheme that explains and generalizes many of the metrics previously applied in the state of the art.

The first component of our framework, the item novelty model, particularizes the novelty and diversity of recommendations to the individual contributions of the items that compose them. Our novelty model is in turn based on a novelty context, which determines the notion of novelty or diversity considered, and a measurement approach, that translates the perception of novelty into a numerical value. We have suggested several alternatives for item novelty models that result in previously proposed metrics of the state of the art and further new metrics. The second component, a browsing model inspired in recent work on the formalization of metrics in Information Retrieval, adds two important properties for the evaluation of recommendation lists, namely rank and relevance awareness, which are absent in most prior approaches for measuring the novelty and diversity of recommendations.

Re-ranking strategies for diversity or novelty enhancement are automatically derived from the proposed framework, based on the direct optimization of item novelty models by means of simple re-scoring or greedy selection over the output of baseline recommendation algorithms. By making a trade-off between the novelty of items and their predicted relevance, our re-ranking strategy permits to control the desired degree of novelty and diversity in the final recommendations, which gets reflected when considering ranking in relevance in the target metrics.

Our experiments validate the proposed approach and provide further observations on the behavior of metric variants – specially when considering the relevance of recommended items – and the effect of the re-ranking strategies. Finally, the results of our framework allow an analysis of the connections and differences between metrics among and within the different perspectives on novelty and diversity.

8.2.2 *Intent-Aware Diversity*

In Chapter 5 we have adapted the family of Intent-Aware metrics and diversification methods of search result diversification to Recommender Systems. We have presented also two new diversification methods under this framework that improve over a direct adaptation of a well-known diversification technique in Information Retrieval.

The adaptation of the Intent-Aware framework for evaluating and enhancing the diversity of search results to recommendations is done by considering a notion of user aspect. We find that the problems that motivate promoting diverse documents in search, namely query ambiguity and underspecification, are analogous to the ambiguity expressed in the heterogeneity of users interests and tastes as expressed in their profiles. We hence propose the concept of user aspect spaces as a proxy for elementary components of users interests in a particular recommendation domain. By replacing query aspects with our user aspects, we adapt several diversity metrics (S-recall, ERR-IA, α-nDCG) and diversification methods (xQuAD, IA-Select) of Information Retrieval to, respectively, evaluate and enhance the diversity of recommendations.

Beyond the direct adaptation of metrics and methods in Information Retrieval, we have devised two new diversification methods that achieve competitive results when compared to the direct adaptations and bring further advantages in terms of additional properties. The first method, an explicit relevance model, replaces the commonly assumed item choice probability found in the considered diversification models by the probability of relevance of the items. In particular, applying such relevance-based model to the xQuAD diversification algorithm of Santos et al. (2010a) results in the relevance-based xQuAD (RxQuAD) algorithm. Our relevance-based approach, besides providing improvements over the direct adaptation of xQuAD, allows moreover a principled manner to control the tolerance for redundancy to adapt to different conditions that influence how redundancy should be handled.

The second method we have proposed combines recommendations targeting particular interests of the user into single, diversified recommendations to properly capture the heterogeneity of user tastes. In particular, we have considered subsets of user profiles covering particular interests, which we call sub-profiles, to better generate recommendations for each specific interest of the user. Once recommendations are generated for each sub-profile, we combine them by means of variants of the xQuAD and RxQuAD method called SxQuAD and SRxQuAD, respectively.

An experimental evaluation shows the validity and soundness of our adaptation of the Intent-Aware framework in recommendation, the performance of our two proposed diversification methods and the redundancy management of the explicit relevance modeling approach.

8.2.3 *Coverage, Redundancy and Size-Awareness in Genre-Based Diversity*

In Chapter 6 we have proposed a recommendation-specific approach to promote the Intra-List Diversity of recommendations. In particular, we have considered the use of genres in domains such as movie, music and book recommendations and have identified properties that genre-based diverse recommendations should fulfill. Addressing these requirements, we have presented a new Binomial framework to measure and optimize genre-based diversity.

Our study starts by analyzing the properties of genres when compared with the subtopics of search result diversification as defined in the TREC Web track. Genre generality and overlap are identified as two characteristics mostly absent in TREC subtopics that condition the direct application of prior methods for search or recommendation list diversification when relying on genre information. Then, based on the previous observations and prior work, we have defined three requirements that genre-based diversity in recommendations should meet: coverage, redundancy and size-awareness. Prior approaches in Information Retrieval diversity do not properly account for these three aspects when considering genres as the source of diversity.

Our proposal to take coverage, redundancy and size-awareness into account is the Binomial framework. This approach is inspired in random recommendation as a natural source of diversity in recommendations. By modeling the occurrence of a genre in a recommendation list as a series of Bernoulli trials, we provide scores to assess the coverage and redundancy for a given recommendation list size according to the generality of genres and the preferences of the user. Additionally, we define a re-ranking strategy to promote genre-based diversity based on greedy selection.

A set of experiments illustrate the properties of our framework and compares it with prior approaches for measuring diversity by means of genres. In particular, our experiments provide confirmation for the limitations of the compared alternatives for adequately addressing coverage and redundancy as opposed to our framework. Furthermore, they illustrate the particular properties of our framework, i. e. the trade-off between genre generality and personalized relevance and the size-awareness of our proposal.

8.2.4 *Recommending Users to Items to Improve Sales Diversity*

Our final contribution in Chapter 7 improves Sales Diversity – and Long Tail Novelty – by inverting the recommendation task by recommending users to items. As previously studied in the area, the popularity bias that collaborative filtering methods suffer can potentially reduce the utility of recommendations for both users and the underlying business. In particular, we have been interested in the effect of such bias in concentrating the recommendations, and thus the potential sales, into a set of reduced, popular items. We have proposed two new methods, inverted nearest neighbors and a probabilistic reformulation, to alleviate the effect of the popularity bias and to improve Sales Diversity and, indirectly, Long Tail Novelty.

Our first method results from the inversion of the recommendation task in collaborative filtering. In fact, by changing the role of users and items in the well-known user and item-based nearest neighbors algorithms, we obtain a new policy for selecting user and item neighborhoods. This new method, which we call inverted nearest neighbors, reduces the concentration of users and items composing the neighborhoods, resulting in more novel and diverse recommendations.

The second proposal takes a probabilistic approach to the recommendation problem, in which a formulation derived from applying the Bayes rule allows us to control the amount of popularity bias in the recommendations.

Experiments in three well-known datasets attest the effectiveness of our two proposals. On one hand, inverted neighborhoods outperform standard neighborhoods in terms of Long Tail Novelty and Sales Diversity – achieving comparable performance to that of latent factor models, which behave better with respect to the popularity bias– while maintaining the same level of accuracy of results. On the other hand, our probabilistic reformulation layer achieves a better trade-off between accuracy and novelty and diversity of results than prior approaches that target the improvement of Long Tail novelty of recommended items.

8.3 FUTURE WORK

The contributions of this thesis introduce new ways of evaluating and enhancing novelty and diversity in Recommender Systems. Though we have covered, to a greater or lesser degree, many angles on the topic, the work presented here opens the way for further extensions. Very particularly, user studies would bring additional validation and further insights to the addressed questions and the proposed approaches, so far examined by theoretical analysis and offline experimentation. Moreover, some of the contributions can be further extended to domains other than Recommender Systems.

We outline next some of the envisioned lines of work to continue the work of this thesis. First, Section 8.3.1 underline the utility of conducting user studies to validate our proposals of assessment and optimization of novelty and diversity in recommendations. Second, we present in Section 8.3.2 possible lines of extension of our contributions. Finally, in Section 8.3.3 we point at domains other than Recommender Systems that might benefit from the contributions presented in this thesis.

8.3.1 *User Studies*

In Chapter 4 we have added rank and relevance sensitivity to the evaluation of the different perspectives on novelty and diversity that influence the utility perceived by users in the recommendations. In Chapter 5 we have considered Intra-List Diversity as a property to cope with the ambiguity of user profiles given by the typical variety of interests and tastes manifested in them. In Chapter 6 we have considered the specific case of genre-based diversity and argued that coverage, redundancy and recommendation list size-awareness are the three basic requirements to model the diversity perceived by users in such setting. As it can be noticed, in all these three chapters a number of assumptions about how users interact with recommendations and perceive their utility are made.

Although such assumptions are based on theories and evidence found in prior work in the areas of Recommender Systems and Information Retrieval, user studies might bring further understanding of the perception of novelty and diversity in recommendations. Such studies should serve two purposes: first to validate the principles that are the basis of our proposals and then help us decide between different parameters and configurations.

Prior work has conducted user studies to evaluate the perception of users of novelty and diversity and their effect in satisfaction with recommendations (Ziegler et al., 2005; Bollen et al., 2010; Pu et al., 2011; Ekstrand et al., 2014). However, we are not aware of any prior work in Recommender Systems in which different models or alternative configurations of one model of novelty or diversity are contrasted with real user feedback to determine the appropriateness of one or another alternative. We consider therefore that our envisioned user studies should put the focus in determining which are the best set of models, assumptions and principles that are consistent with user perception of novelty and diversity in recommendations.

8.3.2 *Extending our Contributions*

Another line of future work consists in the extension of our work by means of further explorations of the possibilities of our proposals. We detail here some of the envisioned extensions of our contributions.

Regarding Chapter 6, we envision the extension of our Binomial framework to measure Unexpectedness, which is the other perspective on novelty and diversity in recommendations for which genres can be used. Another way for the extension of our Binomial framework may consist in applying it to features of the items other than genres but with similar characteristics in terms of coverage and redundancy, such as tags, languages, release date, etc.

In Chapter 7, we envisage deeper studies on the properties of neighborhoods we examined in Section 7.4 with additional metrics in order to uncover further potential biases in user and item neighborhoods. We also envision further improvements of the probabilistic reformulation. In particular, we intend to explore increasing the *exclusivity* of items, that is, recommending each item to only a limited selection of users. Also, we think that the parameterization of the item prior in this reformulation can be used for tasks other than alleviating the popularity bias. For instance, in some situations the business might be interested in promoting certain items as a "hidden agenda" (Azaria et al., 2013), namely, addressing business-exclusive goals. By conveniently defining our item prior to assign higher probabilities for those items, our probabilistic reformulation could be conveniently adapted to this task.

8.3.3 *Application of our Contributions to other Fields*

Finally, we believe that our contributions to novelty and diversity in Recommender Systems can be applied to other domains. As we have reviewed in Section 2.3.2, there are other fields such as Sociology, Psychology, Economy, Ecology, Genetics, Telecommunications or Information Retrieval in which novelty and diversity – in one or other of their interpretations – are considered. We conjecture that, as well as we have borrowed ideas and concepts from these fields – particularly from Information Retrieval –, it would be possible to apply our contributions to these domains. In particular, we suggest in this section the potential usefulness of applying our recommendation-specific contributions in Chapters 6 and 7 to the search task.

In Chapter 6 we have stressed the differences between diversity in search and recommendation and specifically how genres and TREC subtopic are not necessarily interchangeable in all situations. In particular, a specific redundancy management is introduced in our Binomial framework to penalize – rather than neutralize – the excess of items covering a particular genre. While we observe that such redundancy management is not required in search diversity, where we are generally interested in retrieving as many subtopics as possible, we wonder whether the application of the Binomial framework to search result diversification would uncover nuances that have been disregarded so far but could bring benefits to the task. We therefore plan to test our Binomial framework in the TREC Web track diversity task and compare it with well-established approaches for this task to unveil the potential usefulness of our proposal in Information Retrieval.

In Chapter 7 we have addressed the improvement of Sales Diversity as a way of avoiding recommendations concentrated in the most popular items of the catalog, which brings benefits to both the businesses and users. Concurrently, in Information Retrieval the concept of Retrievability (Azzopardi and Vinay, 2008) addresses the capacity of an Information Retrieval system of effectively retrieving the documents it indexes. As Azzopardi and Vinay (2008) state, many collections for Information Retrieval evaluation show a retrieval bias quite similar to the popularity bias in recommendations: in extreme cases, 80% of the documents can be removed from the collection without affecting significantly the efficiency of the system. We can easily see that Sales Diversity and Retrievability are closely related. We conceive that the adaptation of our proposals for improving Sales Diversity could help diminish the retrieval bias found in many Information Retrieval collections. Therefore we plan to apply a similar approach to that of Chapter 7 to improve the Retrievability of documents.

A

RANKSYS: A FRAMEWORK FOR THE EXPERIMENTATION OF NOVELTY AND DIVERSITY IN RECOMMENDER SYSTEMS

A.1 INTRODUCTION

RankSys is a new framework for the implementation and evaluation of recommendation algorithms and techniques that has resulted from the work carried out throughout this thesis. While it is envisioned as a framework for the generic experimentation of recommendation technologies, it is naturally specialized in the evaluation and enhancement of novelty and diversity. RankSys receives its name because it targets explicitly the ranking task problem. We therefore do not consider the case of rating prediction as we consider that it leads to sub-optimal recommendations in terms of user satisfaction and business performance. This decision is reflected in the design of the different core interfaces and components of the framework.

The framework has been programmed with Java 8, which is the most recent version of the popular programming language. We take advantage of many of the new features of the language, such as the use of lambda functions, `Stream`'s and facilities for automatic parallelization of the code. The code is licensed under the GPL V3, which allows the free use, study, distribution and modification of the software as long as derived works are distributed under the same license.

To date, the publicly available version of this framework[1] includes the modules that implement novelty and diversity metrics and re-ranking techniques and the required core components of the framework:

- RankSys-core, which contains the common and auxiliary classes of the framework.
- RankSys-metrics, which contains the interfaces and common components for defining metrics.
- RankSys-diversity, which contains the novelty and diversity metrics and re-ranking strategies.
- RankSys-examples, which provides examples of usage of the previous modules.

[1] https://github.com/saulvargas/RankSys

In the rest of the appendix, we provide a high-level description of the different components of the current release of the software.

A.2 INPUT DATA

In the current version of RankSys we consider two types of input data: interactions between users and items – in the form of ratings, play counts, etc. – and feature information about the items – genres, language, etc. In both cases, the data can be interpreted as pairs of entities – user-item pairs and item-feature pairs, respectively – for which we may have some additional information – such as ratings, play counts or weights. In this section, we provide a description of the interfaces and classes employed to represent the input data of our recommendation platform that are part of the RankSys-core module.

As identified in the previous paragraph, our framework considers three different sets of entities, namely users, items and features, whose pairs define the input data of our recommendation algorithms. For each of these entities, we consider an index-like interface that allows us to keep track of its members. For example, the set of users of our system is accessed by means of classes that implement the following interface `UserIndex`:

```
public interface UserIndex<U> {
    public boolean containsUser(U u);
    public int numUsers();
    public Stream<U> getAllUsers();
}
```

Analogous interfaces have been defined for the sets of items (`ItemIndex`) and features (`FeatureIndex`). In the current version, we do not provide direct implementations of these interfaces, but extend them in our input data interfaces as we describe next.

The interaction data between users and items, which is the main information used in collaborative filtering algorithms, is handled by means of classes implementing the interface `RecommenderData`, which extends the `UserIndex` and `ItemIndex` interfaces and adds methods to access the information regarding the interactions between users and items:

```
public interface RecommenderData<U, I, V> extends UserIndex<U>,
 ItemIndex<I> {
    public int numUsers(I i);
    public int numItems(U u);
    public int numPreferences();
    public Stream<IdValuePair<I, V>> getUserPreferences(U u);
    public Stream<IdValuePair<U, V>> getItemPreferences(I i);
}
```

Note that the type of feedback is left as a generic type V, which allows to consider every type of possible feedback data such as ratings, play counts or series of timestamps. Implementations of this RecommenderData can be backed by in-memory structures or by a database. We provide a simple SimpleRecommenderData class that implements this interface by simply storing the user-item pairs in two different hash tables indexed by user and item, respectively.

The information about item features, which can be used by content-based recommendation algorithms or our novelty and diversity metrics and re-ranking techniques, is managed in our framework by an interface FeatureData similar to RecommenderData, which is defined as follows:

```
public interface FeatureData<I, F, V> extends ItemIndex<I>,
 FeatureIndex<F> {
    Stream<IdValuePair<I, V>> getFeatureItems(final F f);
    Stream<IdValuePair<F, V>> getItemFeatures(final I i);
    int numFeatures(I i);
    int numItems(F f);
}
```

Analogously to the user-item data, we provide a hash table-backed implementation of this interface in the class SimpleFeatureData.

A.3 recommendations

As stated in the introduction, our framework considers ranked lists of items as the natural output of recommendation algorithms. In particular, we require the order or recommendation lists to be determined by decreasing order of a scoring function. We therefore define the class Recommendation that encapsulates the information about the user that receives the recommendation and a list of item-score pairs that compose the recommendation:

```
public class Recommendation<U, I> {
    public Recommendation(U user, List<IdDoublePair<I>> items) {...}
    public U getUser() {...}
    public List<IdDoublePair<I>> getItems() {...}
}
```

In our experiments, recommendations are conveniently stored in files for later access of metrics that evaluate their accuracy, novelty or diversity. For that purpose, we include a RecommendationFormat interface whose implementations specify the format in which recommendations are written in and read from files:

```
public interface RecommendationFormat<U, I> {
    ...
    public Writer<U, I> getWriter(OutputStream out) throws IOException;
    public interface Writer<U, I> extends Closeable {
        public void write(Recommendation<U, I> recommendation) throws
         IOException;
    }
    ...
    public Reader<U, I> getReader(InputStream in) throws IOException;
    public interface Reader<U, I> {
        public Stream<Recommendation<U, I>> readAll() throws
         IOException;
    }
}
```

As it can be observed, the RecommendationFormat interface is in turn composed of two different interfaces for reading and writing that have to be reciprocal, that is, the first needs to be able to read the format used by the second. We provide a SimpleRecommendationFormat class that implements this interface by printing in plain text files sets of recommendations as user-item-score triplets sorted by decreasing score for each user.

A.4 METRICS

The common infrastructure for metrics is defined in the RankSys-metrics module. It consists of two different interfaces for metrics and some common components for defining rank and relevance-awareness in metrics.

We identify two types of metrics that evaluate the output of recommendation algorithms: user-oriented metrics that evaluate the ability of a particular recommendation to satisfy the needs of the user that receives it, and business or system-oriented metrics that evaluate the overall effectiveness of a set of recommendations issued to a community of users. For user-oriented metrics, implementations simply have to comply with the following RecommendationMetric interface:

```
public interface RecommendationMetric<U, I> {
    public double evaluate(Recommendation<U, I> recommendation);
}
```

This simple interface provides a numerical value for each recommendation. Its implementations are expected to be immutable, i.e. calculating the result of a recommendation does not change the internal state of the instance of the class. This allows instances of RecommendationMetric to be called concurrently. As examples, we include in RankSys-metrics implementations of two widely used accuracy metrics for ranking tasks, namely precision and nDCG, which are implemented in the classes `Precision` and `NDCG`, respectively. System-oriented metrics have to implement the following, mutable `SystemMetric` interface:

```
public interface SystemMetric<U, I> {
    public void add(Recommendation<U, I> recommendation);
    public void combine(SystemMetric<U, I> other);
    public double evaluate();
public void reset();
}
```

This interface is considerably different to the `RecommendationMetric` interface and has been designed so that the value of the metric can be computed by means of a *mutable reduction*[2] of the recommendations provided, thus allowing the calculation of a metric for a system in a parallel fashion. As an example, we provide in RankSys-metrics a `AverageRecommendationMetric` class that calculates the average value across users of any instance of `RecommendationMetric`.

Additionally, the module RankSys-metrics provides generic models to take into account the ranking and relevance of the items in recommendations. The ranking model defined in the interface `RankingDiscountModel` defines a discount function based on the rank of an item in a recommendation:

```
public interface RankingDiscountModel {
    public double disc(int k);
}
```

Implementations of this interface can be plugged into metrics to consider different ranking discounts. We provide four different classes that implement this interface: `NoDiscountModel` for ignoring any rank position discount, `LogarithmicDiscountModel` as in nDCG, `ExponentialDiscountModel` as in RBP and `ReciprocalDiscount` as in ERR. The relevance model considers the perception of the users about the relevance of the recommended items. It is defined in the abstract class `RelevanceModel`, which extends an auxiliary `PersonalizableModel` class that allows the caching of the resulting user relevance models:

```
public abstract class RelevanceModel<U, I> extends
 PersonalizableModel<U> {
    ...
    protected abstract UserRelevanceModel<U, I> get(U user);
    ...
```

[2] http://docs.oracle.com/javase/8/docs/api/java/util/stream/package-summary.html#MutableReduction

```
    public interface UserRelevanceModel<U, I> extends UserModel<U> {
        public boolean isRelevant(I item);
        public double gain(I item);
    }
}
```

As it can be observed, the abstract class `RelevanceModel` defines an interface `UserRelevanceModel` that has two methods to determine whether an item is found relevant to a user and the gain that is obtained when recommending it. An extension to this interface is defined in `IdealRelevanceModel` to retrieve the set of all items that the user finds relevant, which is useful in metrics that are normalized by the maximum possible score, such as nDCG. We include in this module two instantiable relevance models: `NoRelevanceModel` to ignore the relevance of items and `BinRelevanceModel` which applies a threshold on a test partition of the user-item interaction data to determine the relevance of the recommended items. Some of the metrics provided in this module or the novelty and diversity metrics in RankSys-diversity may define their own relevance models depending on the definition of the metric they are implementing.

A.5 NOVELTY AND DIVERSITY

The module RankSys-diversity contains our implementations of the novelty and diversity metrics and re-ranking strategies that have been implemented for this thesis. The metrics implement the interfaces defined in the module RankSys-metric, while the re-ranking methods share a common set of interfaces and classes defined in this module. In this section, we provide a description of the common re-ranking interfaces and classes and an overview of the different novelty and diversity models grouped in the different packages of the module.

A.5.1 *Re-ranking Strategies*

In this thesis we have considered the re-ranking of the output of baseline recommendation algorithms as a practical and efficient way of optimizing the novelty and diversity of recommendations. We consider two types of re-ranking: one that is the result of a direct re-scoring of the scores provided by the original recommendation ranking, and a greedy selection in which some set-wise magnitude is maximized by iteratively selection those items that maximize it. In both cases, we consider a high-level interface `Reranker` which, given an original recommendation, returns another recommendation that is a re-ranking of the first:

```
public interface Reranker<U, I> {
    public Recommendation<U, I> rerankRecommendation
     (Recommendation<U, I> recommendation);
}
```

In our implementation, we consider an abstract class PermutationReranker that, rather than returning a Recommendation object, returns the permutation that results from the re-ranking. The purpose of this PermutationReranker is to have a more compact representation of re-rankings. By saving only the permutation that defines the recommendation, we can efficiently store in disk or keep in memory many re-rankings of a single recommendation baseline. As a direct instantiable implementation of this PermutationReranker, we include a RandomReranker which returns randomly generated permutations. Re-ranking strategies based on direct re-scoring of a recommender's output also implement directly this interface. Re-ranking methods based on greedy selection extend the abstract class GreedyReranker, which performs a greedy selection based on an objective function that is updated after each step of the selection. Since most of our greedy re-ranking algorithms are themselves based on a linear combination of the original recommender's scores and some novelty component, we provide an abstract LambdaReranker class that performs a normalized linear combination of the original scoring and the novelty component.

A.5.2 *Item Novelty Metrics and Re-Ranking Strategies*

The user-oriented metrics defined in Chapter 4 – with the exception of EILD – are implemented in package es.uam.eps.ir.ranksys.diversity.itemnovelty. This package includes a generic ItemNovelty interface for personalized novelty models, which is the base for the abstract ItemNoveltyMetric class for metrics and the abstract ItemNoveltyReranker class for direct re-ranking strategies:

```
public abstract class ItemNovelty<U, I> extends
 PersonalizableModel<U> {
    ...
    public UserItemNoveltyModel<U, I> getUserModel(U u) {...}
    public interface UserItemNoveltyModel<U, I> extends UserModel<U> {
        public double novelty(I i);
    }
}
```

In the current version of the framework, three sub-classes of ItemNovelty are included to represent the popularity complement (PC, see Equation 4.2), free discovery (FD, see Equation 4.3) and profile distance (PD, see Equation 4.4) item novelty models defined in Chapter 4.

A.5.3 *Distance-Based Metrics and Re-Ranking Strategies*

For better readability of the code, the intra-list distance-based metrics and re-ranking algorithms of Chapter 4 do not extend from the previous item novelty model package and are separated in its own package `es.uam.eps.ir.ranksys.-diversity.distance`. This package defines an `ItemDistanceModel` for considering different definitions for the distance between items:

```
public interface ItemDistanceModel<I> {
    public double dist(I i, I j);
}
```

We include an abstract class `FeatureItemDistanceModel` that takes a `FeatureData` object to compute the distance between items by means of their features. Two distance functions based on Jaccard and cosine similarity are implemented by extending `FeatureItemDistanceModel`. On top of this distance models, the `EILD` class provides the implementation of the EILD metric and, respectively, the `MMR` class implements the corresponding re-ranking strategy.

A.5.4 *Sales Diversity Metrics*

Rank and relevance-unaware Sales Diversity metrics in Chapter 4 are implemented in the package `es.uam.eps.ir.ranksys.diversity.sales.metrics`. Since these are business-oriented metrics, they implement the interface `SystemMetric`. In particular, the implemented metrics are Aggregate Diversity, Entropy, Gini Index and Gini-Simpson Index. Since the three last metrics are based on the number of times each item is recommended to the community of users, they conveniently extend the abstract `AbstractSalesDiversityMetric` class that implements the count of how many items each item is recommended to users.

A.5.5 *Intent-Aware Metrics and Re-Ranking Strategies*

Our adaptation of the Intent-Aware metrics and diversification techniques in Chapter 5 is contained in the package `es.uam.eps.ir.ranksys.diversity.intentaware`. The basis of this package is the `IntentModel` class, which represents the concept of user aspect space when it is defined by item features in the user profile. This `IntentModel` is then used in the implementations of the metrics ERR-IA and α-nDCG and the xQuAD diversification method provided in this package.

A.5.6 *Binomial Metrics and Re-Ranking Strategies*

The metrics and re-ranking strategies of the Binomial framework proposed in Chapter 6 are found in package `es.uam.eps.ir.ranksys.diversity.binom`. All of them use the `BinomialModel` class, which implements the binomial probability model that defines the coverage and redundancy scores for a given recommendation list size. Metrics and re-ranking strategies for coverage, redundancy and joint diversity are included in this package.

A.6 examples

The module RankSys-examples contains two examples of use of the metrics and re-ranking strategies defined in the other modules. Together with this documentation, they should be used as a starting point to familiarize with the code in the framework. As more modules are added to the framework, additional example code will be added to this module.

B

INTRODUCCIÓN

B.1 MOTIVACIÓN

Los Sistemas de Recomendación se han convertido en una tecnología ubicua en un aplico espectro de aplicaciones del día a día, y se puede decir que son familiares para el público. Desde la creación de la World Wide Web en 1991 (Berners-Lee, 1992), la cantidad de contenido y recursos en esta ha crecido exponencialmente. El caso de las plataformas de comercio electrónico y *streaming* y las redes sociales, las cuales constituyen a día de hoy una sustancial parte del tráfico de la *Web*, es especialmente interesante. Típicamente, estos servicios ofrecen una variada y amplia selección de contenidos a sus clientes: más de 200 millones de productos en Amazon.com, 30 millones de canciones en Spotify, 10.000 películas en Netflix, 248 millones de usuarios activos en Twitter enviando 500 millones de mensajes diarios, etc. En tales situaciones donde hay una *sobrecarga de información*, ayudar a los usuarios a explorar y encontrar recursos de su interés es vital para la viabilidad de esos modelos de negocio. Las tecnologías de recomendación, al proveer sugerencias personalizadas en catálogos masivos, han resultado ser una fuente destacada de ingresos y satisfacción de usuarios.

Los Sistemas de Recomendación han atraído un creciente nivel de interés en la comunidad académica en las últimas dos décadas. Esto ha resultado en una abundancia de algoritmos de recomendación, tecnologías y software. La investigación en el área ha sido cubierta en muchos foros. La ACM Conference on Recommender Systems[1], celebrada desde 2007, se puede considerar el principal foro en el campo. Los Sistemas de Recomendación son un tema recurrente en otras conferencias de alto nivel en Ciencias de la Computación, como son WWW (Sarwar et al., 2001), SIGIR (Herlocker et al., 1999), CIKM (Karatzoglou et al., 2012), WSDM (Zhang et al., 2012), ICML (Salakhutdinov et al., 2007) o KDD (Niemann and Wolpers, 2013), por nombrar unas pocas. Revistas como TKDE (Yu et al., 2004), TOIS (Herlocker et al., 2004), IPM (Sweeney et al., 2008) o IRJ (Wang et al., 2008) son también una fuente principal de publicaciones de investigación en este campo.

A pesar del considerable progreso que se ha hecho en las últimas dos décadas en el área de los Sistemas de Recomendación, hay una percepción general de que hay todavía muchos retos y problemas abiertos que afectan al estado actual de las tecnologías de recomendación y requieren de esfuerzos adicionales. Identificamos

[1] http://recsys.acm.org/

novedad y diversidad en Sistemas de Recomendación, las cuales han atraído la atención de la comunidad en la última década, como dimensiones fundamentales de la calidad en Sistemas de Recomendación cuyo estudio constituye una dirección prometedora para el avance en el campo. Las contribuciones de esta tesis se enmarcan en este tema en particular.

Novedad y diversidad cubren un conjunto de perspectivas diferentes pero interrelacionadas que afectan a la calidad de las recomendaciones en términos de satisfacción de los usuarios y desempeño de negocio. Hay muchas situaciones donde se puede mejorar la satisfacción del usuario por las recomendaciones al aplicar algún grado de novedad o diversidad. Considere el caso de recomendar la *canción del verano*. Asumiendo que el usuario que recibe tal recomendación escucha música con frecuencia, hay una alta probabilidad de que este ya la conozca. En este caso, sugerir canciones más nuevas (en el sentido de menos populares) podría contribuir a la utilidad de las recomendaciones como herramientas de descubrimiento de contenido nuevo y desconocido. En un contexto diferente, recomendar una lista de películas consistente en, pongamos por ejemplo, *westerns*, independientemente de lo que le gusten al usuario, puede ser altamente redundante e insatisfactorio para las necesidades del usuario. Los usuarios tienden a tener una variedad de intereses y gustos y una necesidad de recomendaciones diversas, por tanto centrarse en recomendar un género cinematográfico en particular puede resultar en una recomendación sub-óptima. En términos de desempeño de negocio, también se puede considerar novedad y diversidad. Por ejemplo, ofrecer recomendaciones diferentes a los distintos usuarios tiene sentido desde el punto de vista del negocio: no sólo los usuarios están interesados en explorar el catálogo, sino que el propio negocio está interesado en hacer todo el catálogo visible a sus usuarios. Un sistema que proporciona recomendaciones altamente relevantes usando sólo una décima parte del catálogo puede satisfacer a los usuarios, pero es sub-óptimo desde un punto de vista de negocio. Estas y otras perspectivas de novedad y diversidad son el objeto de esta tesis.

La frase "Si no lo puedes medir, no puedes mejorarlo", atribuida a Lord Kelvin, se aplica perfectamente a la evaluación de los Sistemas de Recomendación. En efecto, evaluar apropiadamente el desempeño de las recomendaciones en términos de las diferentes dimensiones de calidad involucradas es el primer paso para hacerlas útiles. Desde principios de la década pasada, un creciente número de propuestas ha resultado en una variedad de métricas para la evaluación de las diferentes perspectivas de novedad y diversidad en Sistemas de Recomendación. No obstante, encontramos en este conjunto de métricas una heterogeneidad y falta de análisis detallado sobre propiedades tan importantes como considerar la utilidad real proporcionada al usuario en términos de la posición y relevancia de los objetos recomendados: por muy novedoso que sea un objeto recomendado, se obtendrá poca utilidad si al usuario no le gusta o ni siquiera lo ve. Consideramos que necesitamos revisitar el trabajo relacionado con este tema bajo una perspectiva renovada. En particular, encontramos en el grado de formalización en la evaluación de la Recuperación de Información, especialmente en lo relativo a la evaluación de la diversidad de resultados de búsqueda, una fuente prometedora de teorías y conceptos que podría ayudarnos a trazar apropiadamente una nueva visión sobre la evaluación en recomendaciones.

El área de la Recuperación de Información trata la tarea más amplia de proporcionar a los usuarios un acceso fácil a la información de su interés. Trata sobre la representación, almacenamiento, organización y acceso a elementos de información como documentos, páginas web, catálogos en línea, registros estructurados o semi-estructurados y objetos multimedia (Baeza-Yates and Ribeiro-Neto, 2011). Los motores de búsqueda Web son las herramientas de Recuperación de Información más visibles. Dado un usuario que expresa una *necesidad de información* en forma de una consulta (breve), la tarea de un motor de búsqueda Web consiste en devolver un resultado de búsqueda compuesta de páginas Web relevantes con respecto a la consulta efectuada. Los Sistemas de Recomendación se pueden ver como un caso especial de un sistema de Recuperación de Información en el cual la *necesidad de información* se expresa de forma implícita – esto es, no hay consulta – y por tanto la personalización es particularmente decisiva para satisfacer la *necesidad de información* del usuario. Es por tanto natural contemplar puntos de vista comunes entre búsqueda y recomendación y adaptar técnicas de un campo al otro. Una parte importante de nuestras contribuciones resulta de adaptar nociones de Recuperación de Información a los Sistemas de Recomendación.

La evaluación de sistemas de Recuperación de Información se ha caracterizado por la formalización de métricas y metodologías de evaluación bajo conceptos bien comprendidos y modelos de usuario elaborados (p.e. Carterette (2011)). Creemos que ese nivel de rigor en la definición de métricas puede beneficiar a la todavía incipiente evaluación de novedad y diversidad en Sistemas de Recomendación. Además, la diversidad en Recuperación de Información también tiene un papel importante para lidiar con la ambigüedad y sub-especificación de las consultas. Una consulta como "java" puede referirse tanto al lenguaje de programación como a la isla indonesia. En este caso, presentar al usuario documentos que cubran estas u otras posibles interpretaciones es una estrategia efectiva para satisfacer las posibles *necesidades de información* subyacentes a la consulta. En el espíritu de buscar perspectivas comunes entre búsqueda y recomendación, vemos aplicable la motivación para diversificación en búsqueda al caso de recomendación: los usuarios suelen tener una variedad de intereses que, en la ausencia de información adicional, tienen que ser que ser cubiertos en las recomendaciones que reciben. Por tanto, pensamos que explorar la adaptación de la diversificación de resultados de búsqueda a la tarea de recomendación puede traer todavía más beneficios a la evaluación y mejora de novedad y diversidad en Sistemas de Recomendación.

Los Sistemas de Recomendación presentan, sin embargo, particularidades que necesitan ser tratadas con motivaciones y técnicas específicas para este dominio. La diversidad dentro de las recomendaciones no sólo soluciona un potencial problema de ambigüedad de las necesidades de los usuarios, sino que también trata la necesidad o deseo de recibir recomendaciones variadas. Por tanto, se requiere un análisis de las propiedades de recomendaciones diversas para poder ir más allá de lo que ofrecen los últimos avances en recomendación y diversificación de resultados de búsqueda. La diversidad entre las recomendaciones recibidas por distintos usuarios is también un problema específico apenas abordado en Recuperación de Información. En la mayoría de escenarios de recomendación se da el conocido como *efecto de larga cola* (Anderson, 2006), en el cual el pequeño conjunto de los artículos más populares (la cabeza) representa una porción significativa de las interacciones con los usuarios (en forma de visionados, puntuaciones o ventas), frente al resto de los artículos (la larga cola). Diversos autores han sostenido que un sistema de recomendación que promueve recomendaciones en la larga cola no solo tiene beneficios para el negocio en tanto que se aprovecha al máximo el catálogo, sino que proporciona al usuario con recomendaciones menos conocidas u obvias (Celma and Herrera, 2008). En esta tesis, tratamos tales problemas específicos a la recomendación y proponemos soluciones para su mejora.

B.2 OBJETIVOS

El principal propósito de esta tesis es proponer un enfoque fundamentado de la evaluación y mejora de novedad y diversidad en Sistemas de Recomendación. Consideramos que la mejora de tales dimensiones fundamentales de la utilidad de las recomendaciones tiene que tener en cuenta cómo los usuarios exploran y perciben las recomendaciones, cuáles son los problemas que la novedad y la diversidad resuelven, y las causas de los mismos. Para ello, hemos perseguido los siguientes objetivos de investigación:

O1: revisitar el trabajo en evaluación de novedad y diversidad en recomendaciones y desarrollar una base conceptual y metodológica común que tenga en cuenta cómo los usuarios perciben la utilidad de las recomendación. Como se afirma en la motivación, encontramos en el trabajo relacionado sobre la evaluación de novedad y diversidad en Sistemas de Recomendación un amplio conjunto de métricas pero una ausencia de entendimiento claro sobre las conexiones y diferencias entre las perspectivas detrás de las métricas. Aún más, muchas de las métricas carecen de propiedades importantes que reflejan cómo los usuarios interaccionan con las recomendaciones y la utilidad que estas proporcionan.

O2: explorar la aplicación de teorías y métodos de diversidad en Recuperación de Información a los Sistemas de Recomendación. Al vincular las tareas de recomendación y búsqueda, investigamos los beneficios de aplicar técnicas de diversificación en resultados de búsqueda a la tarea de recomendación. En particular, establecemos una analogía entre la ambigüedad y sub-especificación de consultas breves efectuadas en motores de búsqueda y la ambigüedad de intereses y gustos de los usuarios junto con la incertidumbre inherente sobre estos intereses en un sistema de recomendación.

O3: concebir técnicas específicas para recomendación para mejorar la diversidad de recomendaciones. Más allá de unir perspectivas entre Recuperación de Información y Sistemas de Recomendación, es necesario determinar claramente hasta dónde llegan las similitudes entre recomendación y búsqueda. En particular, la necesidad de recibir recomendaciones diversas está fuera del objetivo de la diversificación de resultados de búsqueda, y requiere de un enfoque dedicado en un contexto de recomendación.

O4: proponer nuevas técnicas para paliar el sesgo de popularidad en recomendaciones. Como se menciona en la motivación, una de las principales causas de la falta de novedad y diversidad es el sesgo en los algoritmos de recomendación hacia los artículos más populares. Proponemos nuevos métodos para paliar este efecto a la vez que se mantiene el acierto en las recomendaciones.

B.3 CONTRIBUCIONES

El trabajo llevado a cabo a lo largo de esta tesis se ha traducido en varias contribuciones a la investigación de la novedad y la diversidad en Sistemas de Recomendación, las cuales resumimos a continuación.

En el **Capítulo** 4 proponemos un **marco unificado para la novedad y la diversidad en Sistemas de Recomendación**. Este marco se basa en tres relaciones fundamentales entre los usuarios y los elementos, a saber, el descubrimiento, la relevancia y la elección, y dos componentes configurables como son un modelo de novedad de artículo y un modelo de navegación, que en conjunto definen y generalizan muchas de las métricas para las distintas nociones de la novedad y la diversidad de la literatura previa. Además de un proporcionar una base formal para las métricas propuestas en trabajo relacionado, nuestro marco admite propiedades adicionales tales como sensibilidad a la posición y relevancia en listas de recomendación.

El **Capítulo** 5 propone la **diversificación de recomendaciones por medio de la adaptación de la familia de métricas y métodos *Intent-Aware* de diversificación de resultados de búsqueda**. Para esta adaptación, se establece una analogía entre las interpretaciones o facetas de consultas y los intereses o gustos de usuarios. Esto nos permite adaptar métricas y métodos de diversificación de resultados de búsqueda a la tarea recomendación. En este contexto, se proponen dos nuevos métodos para mejorar la diversidad de las recomendaciones. El primer método introduce un modelo de relevancia explícita que reemplaza el modelo generativo que se encuentra en las técnicas de diversificación adaptados para diversificación de resultados de búsqueda. El segundo método hace uso de sub-perfiles para proporcionar recomendaciones adecuadas a cada gusto o interés de un usuario diferente, que luego se combinan en una única recomendación diversificada.

Estudiamos en el **Capítulo 6** el caso específico de modelado diversidad en recomendación cuando los artículos se clasifican por medio de géneros, como es el caso de las películas, música o libros. Identificamos la cobertura, la redundancia y la sensibilidad al tamaño de la lista como los requisitos para diversidad basada en géneros, y sostenemos que ninguno de los marcos de diversificación anteriores los satisface. Proponemos un nuevo **marco Binomial para modelar la diversidad basada en géneros** que satisface estos requisitos.

Finalmente, el **Capítulo 7** propone dos métodos para paliar el efecto de la concentración sesgada hacia la popularidad en las recomendaciones basadas en filtrado colaborativo. Ambos enfoques resultan de invertir (conceptualmente) la tarea recomendación al **recomendar usuarios a los artículos**. El primer método es una nueva política para seleccionar vecinos en los algoritmos de vecinos próximos. El segundo método desarrolla una reformulación probabilística de la tarea recomendación que aísla y controla el efecto del sesgo de popularidad en las recomendaciones.

B.4 PUBLICACIONES

El trabajo llevado a cabo para la realización de esta tesis ha dado lugar a varias publicaciones en congresos internacionales, revistas, capítulos de libros y otros foros. Listamos a continuación estas publicaciones, ordenándolas por el tipo de publicación y el capítulo con el que están relacionadas.

Publicaciones Relacionadas con el Capítulo 2

Capítulos de libro:

- Castells, P., Hurley, N., and Vargas, S. (in press). Novelty and diversity in recommender systems. In Ricci, F., Rokach, L., and Shapira, B., editors, *Recommender Systems Handbook 2nd Edition*. Springer US

Publicaciones Relacionadas con el Capítulo 4

Artículos largos en conferencias internacionales y revistas:

- Vargas, S. and Castells, P. (2011). Rank and relevance in novelty and diversity metrics for recommender systems. In *Proceedings of the 5th ACM Conference on Recommender Systems*, RecSys '11, pages 109–116, New York, NY, USA. ACM

Artículos de *workshop*, publicaciones nacionales y *posters*:

- Castells, P., Vargas, S., and Wang, J. (2011). Novelty and diversity metrics for recommender systems: Choice, discovery and relevance. In *International Workshop on Diversity in Document Retrieval at the 33rd European Conference on Information Retrieval*, DDR'11

Publicaciones Relacionadas con el Capítulo 5

Artículos largos en conferencias internacionales y revistas:

- Vargas, S., Castells, P., and Vallet, D. (2012a). Explicit relevance models in intent-oriented information retrieval diversification. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '12, pages 75–84, New York, NY, USA. ACM

- Vargas, S. and Castells, P. (2013). Exploiting the diversity of user preferences for recommendation. In *Proceedings of the 10th Conference on Open Research Areas in Information Retrieval*, OAIR '13, pages 129–136, Paris, France. CID

- Vargas, S., Santos, R. L. T., Macdonald, C., and Ounis, I. (2013). Selecting effective expansion terms for diversity. In *Proceedings of the 10th Conference on Open Research Areas in Information Retrieval*, OAIR '13, pages 69–76, Paris, France. CID

Artículos de *workshop*, publicaciones nacionales y *posters*:

- Vargas, S., Castells, P., and Vallet, D. (2011). Intent-oriented diversity in recommender systems. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '11, pages 1211–1212, New York, NY, USA. ACM

- Vargas, S. and Castells, P. (2012). Diversificación en sistemas de recomendación a partir de sub-perfiles de usuario. In *II Congreso Español de Recuperación de Información*, CERI'12

Publicaciones Relacionadas con el Capítulo 6

Artículos largos en conferencias internacionales y revistas:

- Vargas, S. and Castells, P. (2014a). Improving sales diversity by recommending users to items. In *Proceedings of the 8th ACM Conference on Recommender Systems*, RecSys '14, pages 145–152, New York, NY, USA. ACM

Artículos de *workshop*, publicaciones nacionales y *posters*:

- Vargas, S., Castells, P., and Vallet, D. (2012b). On the suitability of intent spaces for IR diversification. In *Proceedings of the International Workshop on Diversity in Document Retrieval at the 5th ACM International Conference on Web Search and Data Mining*, DDR'12, Seattle, Washington, USA

Publicaciones Relacionadas con el Capítulo 7

Artículos largos en conferencias internacionales y revistas:

- Vargas, S., Baltrunas, L., Karatzoglou, A., and Castells, P. (2014). Coverage, redundancy and size-awareness in genre diversity for recommender systems. In *Proceedings of the 8th ACM Conference on Recommender Systems*, RecSys '14, pages 209–216, New York, NY, USA. ACM

Artículos de *workshop*, publicaciones nacionales y *posters*:

- Vargas, S. and Castells, P. (2014b). Vecindarios inversos para la mejora de la novedad en filtrado colaborativo. In *III Congreso Español de Recuperación de Información*, CERI'14

Otras Publicaciones Relacionadas con la Tesis

Presentación de esta tesis en simposios y consorcios doctorales especializados:

- Vargas, S. (2011). New approaches to diversity and novelty in recommender systems. In *Proceedings of the 4th BCS-IRSG Conference on Future Directions in Information Access*, FDIA'11, pages 8–13, Swinton, UK. British Computer Society

- Vargas, S. (2014). Novelty and diversity enhancement and evaluation in recommender systems and information retrieval. In *Proceedings of the 37th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '14, pages 1281–1281, New York, NY, USA. ACM

B.5 ESTRUCTURA DE LA TESIS

La tesis está estructurada del siguiente modo:

- El Capítulo 1 presenta la motivación, objetivos, contribuciones y publicaciones relacionadas con la tesis y las definiciones y notación.

- El Capítulo 2 repasa el trabajo relacionado en los temas de interés de esta tesis. Primero, proporcionamos un resumen general del trabajo previo en Sistemas de Recomendación. Segundo, nos adentramos en el estudio de novedad y diversidad en Sistemas de Recomendación. Por último, examinamos el trabajo en diversificación de resultados de búsqueda en Recuperación de Información.

- El Capítulo 3 presenta el diseño de los experimentos de nuestras contribuciones. En particular, proporcionamos una descripción detallada de los conjuntos de datos, los algoritmos de recomendación y la metodología de evaluación que hemos usado en las diferentes secciones de experimentos de nuestras contribuciones en los siguientes capítulos.

- En el Capítulo 4 proponemos un marco unificado para novedad y diversidad en Sistemas de Recomendación que contribuye a la formalización de métricas y técnicas de re-*ranking* y la inclusión de propiedades como posición y relevancia en recomendaciones.

- El Capítulo 5 presenta una adaptación de las métricas y métodos de diversificación *Intent-Aware* de diversificación de resultados de búsqueda a los Sistemas de Recomendación. Dentro de esta adaptación, proponemos dos métodos nuevos que mejoran a adaptaciones directas de métodos de diversificación para resultados de búsqueda.

- El Capítulo 6 trata el problema de medir y optimizar la diversidad de las recomendaciones usando géneros. Se identifican tres requisitos para diversidad basada en géneros: cobertura, redundancia y sensibilidad al tamaño de la lista de recomendación. Proponemos un marco Binomial que satisface estos tres requisitos y lo comparamos con otros enfoques relacionados para medir la diversidad de recomendaciones.

- El Capítulo 7 presente nuestras contribuciones para mejorar la diversidad de ventas. Al recomendar conceptualmente usuarios a artículos, presentamos dos propuestas, a saber, vecinos próximos inversos y una reformulación probabilística, que ofrecen mejoras significativas en términos de la diversidad de ventas cuando se comparan con propuestas previas.

- El Capítulo 8 ofrece las conclusiones y el trabajo futuro.

- El Apéndice A contiene la documentación de alto nivel para RankSys, un nuevo *framework* de recomendación desarrollado para esta tesis que se especializa en evaluación y mejora de novedad y diversidad en Sistemas de Recomendación.
- El Apéndice B contiene la traducción al español del Capítulo 1.
- El Apéndice C contiene la traducción al español del Capítulo 8.

B.6 DEFINICIONES Y NOTACIÓN

Resumimos aquí para conveniencia del lector las principales definiciones y notación que utilizaremos en el resto de esta tesis. Otra notación específica, cuando sea necesaria, será descrita en el capítulo donde se aplica.

Sin perdida de generalidad, el problema recomendación puede formularse como sugerir artículos de un catálogo $\mathcal{I}$ en un catálogo de recomendación en particular – productos, películas, música u otros recursos – a una comunidad de usuarios $\mathcal{U}$. Con el fin de hacer recomendaciones personalizadas, requerimos algún conocimiento previo sobre los usuarios en forma de artículos que han puntuado, consumido, comprado, etc. Estos datos de interacción, que denotamos como $\mathcal{R}$, por lo general se codifican en forma de una matriz $|\mathcal{U}| \times |\mathcal{I}|$ cuyos elementos r_{ui} representan la interacción entre los usuarios y los elementos. En su forma más simple, podemos considerar que esta matriz de interacción toma los valores 0 y 1, es decir, $\mathcal{R} \in \{0, 1\}^{|\mathcal{U}| \times |\mathcal{I}|}$, cuando sólo tenemos interacciones binarias entre los usuarios y los artículos: el usuario compró, consumido, vio o escuchó el artículo, etc. En el caso en que los usuarios proporcionan calificaciones numéricas, por ejemplo de 1 a 5 estrellas, $\mathcal{R}$ toma valores en $\{0, 1, 2, 3, 4, 5\}$ donde por 0 denotamos la ausencia de una calificación. En los lugares donde podemos cuantificar la interacción entre los usuarios y los elementos, por ejemplo el recuento de reproducción de música, nuestra matriz de interacción puede tomar valores en el conjunto de los números naturales $\mathbb{N}$. Los casos más complicados, tales como datos de tiempo para las interacciones, pueden ser fácilmente acomodados con formulaciones similares. Convenientemente, abusando de notación podemos re-interpretar, en todos los casos anteriores, la matriz de interacción $\mathcal{R}$ como el conjunto de pares de usuarios y artículos $\mathcal{R} \subset \mathcal{U} \times \mathcal{I}$ que han tenido algún tipo de interacción. Bajo esta interpretación, denotamos como $\mathcal{I}_u = \{i \in \mathcal{I} : (u, i) \in \mathcal{R}\}$ el perfil de usuario o subconjunto de artículos con lo que el usuario u interactuó y, respectivamente, como $\mathcal{U}_i = \{u \in \mathcal{U} : (u, i) \in \mathcal{R}\}$ el perfil de artículo o subconjunto de usuarios que sabemos que tuvieron interacción con el artículo i.

Usando los datos de interacción $\mathcal{R}$, el objetivo de un sistema de recomendación S consiste en generar recomendaciones R_u^S para cada usuario $u \in \mathcal{U}$. Una recomendación es un conjunto de artículos $R_u^S \subset \mathcal{I}$ que se le presentan al usuario.

Frecuentemente, las recomendaciones se presentan como una lista ordenada, así que las podemos interpretar como secuencias de artículos $R_u^S \in \mathcal{I} \times \ldots \times \mathcal{I}$. En nuestra configuración, los artículos recomendados se seleccionan mediante una función de puntuación $s : \mathcal{U} \times \mathcal{I} \rightarrow \mathbb{R}$ cuyo orden decreciente de puntuaciones determina el orden de las recomendaciones., esto es, si $s(u,i) \geqslant s(u,j)$, entonces $R_u^S = (\ldots, i, \ldots, j, \ldots)$. En el resto del documento omitiremos por conveniencia el usuario o el sistema en la notación de recomendación R_u^S y la dejaremos simplemente como R a menos que necesitemos indicar explícitamente uno u otro.

Junto con los datos de interacción $\mathcal{R}$, algunas métricas pueden hacer uso de información adicional acerca de los artículos en la evaluación de algunas perspectivas de novedad y diversidad. Genéricamente, denotaremos como $\mathcal{F}$ al conjunto de rasgos o características sobre los artículos, y como $\mathcal{F}_i$ el subconjunto de características que describen al artículo i. Ejemplos de tales características dependen del tipo de los artículos. Por ejemplo, en el caso de películas o canciones, podríamos considerar el idioma, año de lanzamiento, género, etc. Como se describe en el capítulo 3, consideramos en nuestro diseño experimental el caso concreto de géneros de películas y música. En ese caso, denotamos como $\mathcal{G}$ al conjunto de géneros en un dominio de recomendación específico y, para artículo i, denotamos como $\mathtt{genres}_i$ los géneros cubiertos por este.

C

CONCLUSIONES Y TRABAJO FUTURO

C.1 INTRODUCCIÓN

En esta tesis hemos desarrollado un enfoque fundamentado de la evaluación y mejora de novedad y diversidad en recomendaciones. Primero hemos propuesto un marco unificado para explicar y modelar distintas perspectivas de novedad y diversidad en recomendación. Seguidamente se ha hecho una adaptación de métricas de diversidad y métodos de diversificación de Recuperación de Información a Sistemas de Recomendación, específicamente las métricas y métodos conocidos como *Intent-Aware*. Aún más, hemos presentado una propuesta específica a recomendaciones para medir diversidad de recomendaciones basada en géneros. Por último, hemos afrontado el problema del sesgo de popularidad en recomendaciones y su efecto en la Diversidad de Ventas al desarrollar dos enfoques basados en la inversión parcial de la tarea de recomendación al seleccionar usuarios para artículos en lugar de lo contrario.

En este capítulo presentamos las conclusiones principales que se derivan de nuestras contribuciones y algunas perspectivas de potencial trabajo futuro. En la Sección C.2 presentamos un resumen de este trabajo y las conclusiones. En la Sección C.3 detallamos las vías de investigación futura para extender nuestro trabajo.

C.2 RESUMEN Y CONTRIBUCIONES

Resumimos y destacamos los hallazgos y las contribuciones principales de esta tesis con respecto a los objetivos de investigación formulados en el Capítulo 1.

C.2.1 *Un Marco Unificado para Novedad y Diversidad*

En el Capítulo 4 hemos propuesto un marco unificado para evaluar novedad y diversidad en Sistemas de Recomendación. Al modelar la novedad y diversidad de listas de recomendación mediante un modelo de novedad de artículo y un modelo de navegación, proporcionamos un esquema común que explica y generaliza muchas de las métricas previamente aplicadas en trabajo previo en el área.

El primer componente de nuestro marco, el modelo de novedad de artículo, particulariza la novedad y diversidad de las recomendaciones a partir de las contribuciones individuales de los artículos que las componen. Nuestro modelo de novedad se basa a su vez en un contexto de novedad, que determina la noción de novedad o diversidad considerada, y una estrategia de medición, que traduce la percepción de novedad a un valor numérico. Hemos sugerido distintas alternativas de modelos de novedad que resultan en métricas previamente propuestas en el trabajo previo en el área y otras métricas nuevas. El segundo componente, un modelo de navegación inspirado en el trabajo reciente de formalización de métricas de Recuperación de Información, añade dos propiedades importantes para la evaluación de listas de recomendación, a saber, sensibilidad a la posición y a la relevancia, las cuales no están presentes en la mayoría de las propuestas previas para medir novedad y diversidad en recomendaciones.

Del marco propuesto se derivan automáticamente estrategias de re-*ranking* para la mejora de novedad o diversidad, las cuales están basadas bien en la optimización directa de los modelos de novedad de artículo mediante re-puntuación simple o bien en la selección avara sobre la salida de algoritmos de recomendación de referencia. Al balancear la novedad de los artículos y su predicción de relevancia, nuestra estrategia de re-*ranking* permite controlar el grado deseado de novedad y diversidad en las recomendaciones finales, lo que se refleja cuando consideramos posición y relevancia en las métricas objetivo.

Nuestros experimentos validan nuestra propuesta y proporcionan observaciones adicionales sobre el comportamiento de las variantes de nuestras métricas – especialmente cuando se considera la relevancia de los artículos recomendados – y el efecto de las estrategias de re-*ranking*. Finalmente, los resultados de nuestro marco permiten un análisis de las conexiones y diferencias entre métricas dentro y entre las diferentes perspectivas de novedad y diversidad.

C.2.2 *Diversidad Intent-Aware*

En el Capítulo 5 hemos adaptado la familia de las métricas y métodos de diversificación *Intent-Aware* usadas en diversificación de resultados de búsqueda a Sistemas de Recomendación. Hemos presentado también dos nuevos métodos de diversificación en este marco que mejoran a una adaptación directa de una técnica de diversificación muy conocida en Recuperación de Información.

La adaptación del marco *Intent-Aware* para la evaluación y mejora de la diversidad de resultados de búsqueda a recomendaciones al considerar el concepto de *aspecto de usuario*. Hemos encontrado que los problemas que motivan la diversificación de resultados de búsqueda, a saber, la ambigüedad y la sub-especificación de las consultas, son análogos a la ambigüedad se expresa en la heterogeneidad de intereses y gustos de los usuarios expresada en sus perfiles. Proponemos por tanto el concepto de *espacios de aspectos de usuario* como representantes de los componentes elementales de los intereses de los usuarios en un dominio de recomendación particular. Mediante la sustitución de los aspectos de la consulta con nuestros aspectos de usuarios, adaptamos varias métricas de diversidad (S-recall, ERR-IA, α-nDCG) y métodos de diversificación (xQuAD, IA-Select) de Recuperación de Información para, respectivamente, evaluar y mejorar la diversidad de las recomendaciones.

Aparte de la adaptación directa de métricas y métodos de Recuperación de Información, hemos diseñado dos nuevos métodos de diversificación que consiguen resultados competitivos en comparación con las adaptaciones directas y conllevan otras ventajas en términos de propiedades adicionales. El primer método, un modelo de relevancia explícita, sustituye la comúnmente asumida probabilidad de elección encontrada en los modelos de diversificación por la probabilidad de relevancia de los artículos. En particular, la aplicación de este modelo basado en relevancia al algoritmo de diversificación xQuAD de Santos et al. (2010a) resulta en el nuevo algoritmo xQuAD basado en relevancia (RxQuAD). Nuestro enfoque basado en la relevancia, además de proporcionar mejoras sobre la adaptación directa de xQuAD, permite además controlar de manera fundamentada la tolerancia a la redundancia para adaptarse a las diferentes condiciones que influyen en cómo se debe manejar ésta.

El segundo método que hemos propuesto combina recomendaciones dirigidas a los intereses particulares del usuario en recomendaciones individuales pero diversas para capturar adecuadamente la heterogeneidad de los gustos de los usuarios. En particular, hemos considerado los subconjuntos de los perfiles de usuario que cubren intereses particulares, lo que llamamos sub-perfiles, para generar mejores recomendaciones para cada faceta del usuario. Una vez que se generan recomendaciones para cada sub-perfil, los combinamos por medio de variantes de los métodos xQuAD y RxQuAD, llamadas SxQuAD y SRxQuAD, respectivamente.

Una evaluación experimental demuestra la validez y la solidez de nuestra adaptación del marco *Intent-Aware* en recomendación, el desempeño de nuestros dos métodos de diversificación propuestos y la gestión de la redundancia del modelo basado en relevancia explícita.

c.2.3 *Cobertura, Redundancia y Sensibilidad al Tamaño en Diversidad Basada en Géneros*

En el Capítulo 6 hemos propuesto un enfoque específico a la tarea de recomendación para promover la Diversidad de Listas. En particular, se ha considerado el uso de géneros en dominios tales como recomendación de películas, música y libros. Se han identificado las propiedades que las recomendaciones diversas basadas en géneros deben satisfacer. Para responder a estos requisitos, hemos presentado un nuevo marco Binomial para medir y optimizar la diversidad basada en géneros.

Nuestro estudio comienza analizando las propiedades de los géneros en comparación con los *subtopics* empleados en diversificación de resultados de búsqueda definidos en el *Web track* de TREC. La generalidad y el solapamiento de géneros son identificados como dos características mayormente ausentes en los *subtopics* de TREC, que condicionan la aplicación directa de los métodos previamente propuestos para diversidad en resultados de búsqueda y recomendaciones usando géneros. Por tanto, basándonos en estas observaciones y el trabajo previo en el área, hemos definido tres requisitos que la diversidad en recomendaciones basada en géneros debe cumplir: cobertura, redundancia y sensibilidad al tamaño. Las propuestas previas en diversidad en Recuperación de Información no representan sin embargo estos tres aspectos cuando se considera géneros como fuente de diversidad.

Nuestra propuesta para tener en cuenta cobertura, redundancia y el sensibilidad al tamaño simultáneamente se denomina marco Binomial. Este enfoque se inspira en la recomendación aleatoria como una fuente natural de la diversidad en las recomendaciones. Al modelar la ocurrencia de un género en una lista de recomendaciones como una serie de ensayos de Bernoulli, proporcionamos mediciones para evaluar la cobertura y redundancia para un tamaño de la lista de recomendaciones dado de acuerdo a la generalidad de los géneros y las preferencias del usuario. Además, también definimos una estrategia de re-*ranking* para promover la diversidad basada en géneros basado en selección avara.

Una serie de experimentos ilustran las propiedades de nuestro marco Binomial y lo comparan con los enfoques anteriores para medir la diversidad basada en géneros. En particular, nuestros experimentos proporcionan una confirmación de las limitaciones de las alternativas comparadas para capturar adecuadamente la cobertura y la redundancia, a diferencia de nuestra propuesta. Además, ilustramos las propiedades particulares de nuestro marco, esto es, el balance entre generalidad y relevancia de géneros y la sensibilidad al tamaño de nuestra propuesta.

C.2.4 *Recomendar Usuarios a Artículos para Mejorar la Diversidad de Ventas*

Nuestra última contribución en el Capítulo 7 mejora la Diversidad de Ventas – y la Novedad *Long Tail* – invirtiendo la tarea recomendación al recomendar usuarios a los artículos. Como se ha estudiado previamente en el área de Sistemas de Recomendación, el sesgo de popularidad que los métodos de filtrado colaborativo sufren puede reducir potencialmente la utilidad de las recomendaciones para los usuarios y el negocio subyacente. En particular, nos hemos interesado por el efecto de tal sesgo en la concentración de las recomendaciones, y por lo tanto el potencial de ventas, en conjuntos reducidos de artículos populares. Hemos propuesto dos nuevos métodos, vecinos próximos invertidos y una reformulación probabilística, para aliviar el efecto del sesgo de popularidad y mejorar la diversidad de ventas e, indirectamente, de la Novedad *Long Tail*.

Nuestro primer método resulta de la inversión de la tarea de recomendación en filtrado colaborativo. De hecho, al intercambiar el papel de los usuarios y los artículos en los algoritmos de vecinos próximos basados en usuario o artículo, obtenemos una nueva política de selección para la selección de los vecindarios de usuario y artículo. Este nuevo método, que llamamos vecindarios próximos invertidos, reduce la concentración de los usuarios y los artículos que componen los vecindarios, lo que resulta en recomendaciones más novedosas y diversas.

La segunda propuesta aplica un enfoque probabilístico al problema recomendación, en el que una formulación derivada de la aplicación de la regla de Bayes nos permite controlar la cantidad de sesgo de popularidad en las recomendaciones.

Los experimentos en tres conjuntos de datos bien conocidos atestiguan la eficacia de nuestras dos propuestas. Por un lado, nuestros vecindarios invertidos superan a los vecindarios estándar en cuanto a la Novedad *Long Tail* y Diversidad de Ventas – consiguiendo resultados comparables a los de los modelos de factores latentes, que se comportan mejor con respecto al sesgo de popularidad – a la vez que se mantiene el mismo nivel de precisión de los resultados. Por otro lado, nuestra capa de reformulación probabilística obtiene un mejor balance entre la precisión, la novedad y la diversidad que los resultados de enfoques que optimizan directamente la Novedad *Long Tail* de los artículos recomendados.

C.3 TRABAJO FUTURO

Las contribuciones de esta tesis presentan nuevas formas de evaluar y mejorar la novedad y la diversidad en Sistemas de Recomendación. Aunque hemos cubierto, en mayor o menor grado, muchos puntos de vista sobre el tema, el trabajo que aquí se presenta abre el camino a extensiones adicionales. Muy en particular, creemos que la realización de estudios de usuario traería una validación y conocimiento adicionales a las preguntas abordadas y a las propuestas realizadas, hasta ahora examinadas mediante análisis teóricos y experimentación *offline*. Aún más, algunas de las contribuciones se puede ampliar o, incluso, aplicar a dominios distintos a los Sistemas de Recomendación.

Describimos a continuación algunas de las líneas de trabajo previstas para continuar el trabajo de esta tesis. En primer lugar, la Sección C.3.1 subraya la utilidad de llevar a cabo estudios de usuario para validar nuestras propuestas de evaluación y optimización de novedad y diversidad en las recomendaciones. En segundo lugar, se presentan en la Sección C.3.2 posibles líneas de extensión de nuestras contribuciones. Por último, en la Sección C.3.3 señalamos dominios distintos a los Sistemas de Recomendación que podrían beneficiarse de las contribuciones presentadas en esta tesis.

C.3.1 *Estudios de Usuario*

En el Capítulo 4 hemos añadido sensibilidad a la posición y a la relevancia para la evaluación de las diferentes perspectivas de novedad y diversidad que influyen en la utilidad percibida por los usuarios en las recomendaciones. En el Capítulo 5 hemos considerado la Diversidad de Listas como una propiedad para hacer frente a la ambigüedad de los perfiles de usuario provocada por la típica variedad de intereses y gustos de los usuarios. En el Capítulo 6 hemos considerado el caso específico de la diversidad basada en géneros y argumentamos que la cobertura, la redundancia y la sensibilidad al tamaño de la recomendación son los tres requisitos básicos para modelar la diversidad percibida por usuarios. Como se puede apreciar, en todos estos tres capítulos se hacen una serie de hipótesis sobre cómo los usuarios interactúan con las recomendaciones y perciben la utilidad de las mismas.

Aunque estas suposiciones se basan en teorías y evidencias encontradas en el trabajo previo en las áreas de Sistemas de Recomendación y Recuperación de Información, los estudios de usuario podrían ofrecer un mayor entendimiento de la percepción de la novedad y la diversidad en las recomendaciones. Tales estudios deben servir dos propósitos: primero para validar los principios que son la base de nuestras propuestas y luego para ayudarnos a decidir entre los diferentes parámetros y configuraciones de estos.

Trabajos anteriores han llevado a cabo estudios de usuario para evaluar la percepción de los usuarios de la novedad y la diversidad y sus efectos en la satisfacción con las recomendaciones (Ziegler et al., 2005; Bollen et al., 2010; Pu et al., 2011; Ekstrand et al., 2014). Sin embargo, no tenemos conocimiento de ningún trabajo previo en Sistemas de Recomendación en el que diferentes modelos o configuraciones alternativas de un modelo de novedad o diversidad se contrasten con la percepción de usuarios reales para determinar la idoneidad de uno u otro. Consideramos, por tanto, que nuestros estudios de usuario previstos deben poner el foco en la determinación de cuáles son los mejores conjuntos de modelos, suposiciones y principios consistentes con la percepción del usuario de novedad y diversidad en las recomendaciones.

C.3.2 *Extensión de Nuestras Contribuciones*

Otra línea de trabajo futuro consiste en la ampliación de nuestro trabajo por medio de la exploración de las nuevas posibilidades de nuestras propuestas. Detallamos aquí algunas de las extensiones previstas de nuestras contribuciones.

Respecto al Capítulo 6, prevemos la extensión de nuestro marco Binomial para medir Sorpresa, que es otra de las perspectivas de novedad y diversidad en recomendaciones para la que hemos utilizado géneros. Otro camino para la extensión de nuestro marco Binomial puede consistir en su aplicación a otras características de los artículos aparte de géneros pero con propiedades similares en términos de cobertura y redundancia, como pueden ser las etiquetas, idiomas, año de publicación, etc.

En el Capítulo 7, contemplamos estudios más profundos sobre las propiedades de los vecindarios que hemos examinado en la Sección 7.4 con métricas adicionales con el fin de descubrir otros posibles sesgos en los vecindarios de usuario y artículo. También prevemos nuevas mejoras en la reformulación probabilística. En particular, tenemos la intención de explorar el aumento de la *exclusividad* de artículos, es decir, la recomendación de cada artículo a sólo una selección limitada de usuarios. Además, creemos que la parametrización de la probabilidad *a-priori* de los artículos en nuestra reformulación se puede utilizar para tareas distintas de reducir el sesgo de popularidad. Por ejemplo, en algunas situaciones, el negocio detrás de la recomendación se puede interesar en la promoción de ciertos artículos como parte de una "agenda oculta" (Azaria et al., 2013), esto es, abordando objetivos exclusivos al negocio. Definiendo nuestra probabilidad *a-priori* sobre los artículos de tal modo que asigne probabilidades altas a los artículos de interés, creemos que nuestra reformulación probabilística puede ser convenientemente adaptada a esta tarea.

C.3.3 *Aplicación de Nuestras Contribuciones a Otros Campos*

Por último, creemos que nuestras contribuciones a la novedad y la diversidad en Sistemas de Recomendación se pueden aplicar a otros dominios. Como hemos revisado en la Sección 2.3.2, hay otros campos como la Sociología, Psicología, Economía, Ecología, Genética, Telecomunicaciones o Recuperación de Información en los que se consideran la novedad y la diversidad – en una u otra interpretación. Conjeturamos que, al igual que tenemos ideas y conceptos tomados de estos campos – especialmente de Recuperación de Información –, sería posible aplicar nuestras contribuciones a estos dominios. En particular, sugerimos en esta sección la posible utilidad de la aplicación de nuestras contribuciones específicas a recomendación en los Capítulos 6 y 7 para la tarea de búsqueda.

En el Capítulo 6 hemos subrayado las diferencias entre la propiedades de la diversidad en búsqueda y en recomendación y, específicamente, cómo géneros y *subtopics* de TREC no son necesariamente intercambiables en toda situación. Concretamente, una gestión específica de redundancia se introduce en nuestro marco Binomial para penalizar – en lugar de neutralizar – el exceso de artículos que cubren un género particular. Si bien se observa que tal gestión de la redundancia no es necesaria en la diversidad en búsqueda, donde generalmente estamos interesados en recuperar la mayor cantidad *subtopics* como sea posible, nos preguntamos si la aplicación del marco Binomial a diversificación de resultados de búsqueda podría revelar matices que han sido hasta ahora ignorados pero que podrían ser beneficiosos para la tarea. Por lo tanto, tenemos la intención de probar nuestro marco Binomial en la tarea de diversidad del *Web track* de TREC y compararla con el resto de propuestas para esta tarea con el fin de descubrir la utilidad potencial de nuestra propuesta en Recuperación de Información.

En el Capítulo 7 hemos abordado la mejora de la Diversidad de Ventas como una forma de evitar que las recomendaciones estén concentradas alrededor de los artículos más populares del catálogo, lo que trae beneficios para negocios y usuarios. Al mismo tiempo, en Recuperación de Información el concepto de Recuperabilidad (Azzopardi and Vinay, 2008) se refiere a la capacidad de los motores de búsqueda de recuperar los documentos que indexan. Como Azzopardi and Vinay (2008) afirman, muchas colecciones para la evaluación de Recuperación de Información muestran un sesgo de recuperación bastante similar a la sesgo popularidad en las recomendaciones: en casos extremos, hasta el 80% de los documentos de una colección se puede eliminar sin afectar significativamente la eficiencia del sistema. Podemos ver fácilmente que la Diversidad de Ventas y la Recuperabilidad están estrechamente relacionadas. Concebimos que la adaptación de nuestras propuestas para la mejora de la Diversidad de Ventas podría ayudar a disminuir el sesgo de recuperación que se encuentra en muchas colecciones de Recuperación de información. Por lo tanto, tenemos la intención de aplicar un enfoque similar al del Capítulo 7 para mejorar la Recuperabilidad de los documentos.

BIBLIOGRAPHY

Adamopoulos, P. and Tuzhilin, A. (2014). On over-specialization and concentration bias of recommendations: Probabilistic neighborhood selection in collaborative filtering systems. In *Proceedings of the 8th ACM Conference on Recommender Systems*, RecSys '14, pages 153–160, New York, NY, USA. ACM.

Adamopoulos, P. and Tuzhilin, A. (in press). On unexpectedness in recommender systems: Or how to expect the unexpected. *ACM Transactions on Intelligent Systems and Technology, Special Issue on Novelty and Diversity in Recommender Systems*.

Adomavicius, G. and Kwon, Y. (2007). New recommendation techniques for multicriteria rating systems. *IEEE Intelligent Systems*, 22(3):48–55.

Adomavicius, G. and Kwon, Y. (2011). Maximizing aggregate recommendation diversity: A graph-theoretic approach. In *Proceedings of the 1st ACM RecSys Workshop on Novelty and Diversity in Recommender Systems*, DiveRS 2011, pages 3–10.

Adomavicius, G. and Kwon, Y. (2012). Improving aggregate recommendation diversity using ranking-based techniques. *IEEE Transactions on Knowledge and Data Engineering*, 24(5):896–911.

Adomavicius, G. and Kwon, Y. (2014). Optimization-based approaches for maximizing aggregate recommendation diversity. *INFORMS Journal on Computing*, 26(2):351–369.

Adomavicius, G. and Tuzhilin, A. (2005). Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE Transactions on Knowledge and Data Engineering*, 17(6):734–749.

Adomavicius, G. and Tuzhilin, A. (2010). Context-aware recommender systems. In *Recommender Systems Handbook*, pages 217–250. Springer.

Agrawal, R., Gollapudi, S., Halverson, A., and Ieong, S. (2009). Diversifying search results. In *Proceedings of the 2nd ACM International Conference on Web Search and Data Mining*, WSDM '09, pages 5–14, New York, NY, USA. ACM.

Aiolli, F. (2013). Efficient top-N recommendation for very large scale binary rated datasets. In *Proceedings of the 7th ACM Conference on Recommender Systems*, RecSys '13, pages 273–280, New York, NY, USA. ACM.

Amatriain, X. (2012). Building industrial-scale real-world recommender systems. In *Proceedings of the 6th ACM Conference on Recommender Systems*, RecSys '12, pages 7–8, New York, NY, USA. ACM.

Amatriain, X. (2013). Mining large streams of user data for personalized recommendations. *SIGKDD Exploration Newsletter*, 14(2):37–48.

Amigó, E., Gonzalo, J., and Verdejo, F. (2013). A general evaluation measure for document organization tasks. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '13, pages 643–652, New York, NY, USA. ACM.

Anderson, C. (2006). *The Long Tail: Why the Future of Business Is Selling Less of More*. Hyperion.

Azaria, A., Hassidim, A., Kraus, S., Eshkol, A., Weintraub, O., and Netanely, I. (2013). Movie recommender system for profit maximization. In *Proceedings of the 7th ACM Conference on Recommender Systems*, RecSys '13, pages 121–128, New York, NY, USA. ACM.

Azzopardi, L. and Vinay, V. (2008). Retrievability: An evaluation measure for higher order information access tasks. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, CIKM '08, pages 561–570, New York, NY, USA. ACM.

Baeza-Yates, R. A. and Ribeiro-Neto, B. A. (2011). *Modern Information Retrieval - the concepts and technology behind search 2nd edition*. Pearson Education Ltd., Harlow, England.

Balabanović, M. and Shoham, Y. (1997). Fab: Content-based, collaborative recommendation. *Communications of the ACM*, 40(3):66–72.

Belém, F., Santos, R., Almeida, J., and Gonçalves, M. (2013). Topic diversity in tag recommendation. In *Proceedings of the 7th ACM Conference on Recommender Systems*, RecSys '13, pages 141–148, New York, NY, USA. ACM.

Bellogín, A., Cantador, I., and Castells, P. (2010). A study of heterogeneity in recommendations for a social music service. In *Proceedings of the 1st International Workshop on Information Heterogeneity and Fusion in Recommender Systems*, HetRec '10, pages 1–8, New York, NY, USA. ACM.

Bellogín, A., Cantador, I., and Castells, P. (2013). A comparative study of heterogeneous item recommendations in social systems. *Information Sciences*, 221:142–169.

Bellogín, A., Castells, P., and Cantador, I. (2011). Precision-oriented evaluation of recommender systems: An algorithmic comparison. In *Proceedings of the 5th*

ACM Conference on Recommender Systems, RecSys '11, pages 333–336, New York, NY, USA. ACM.

Berners-Lee, T. (1992). The world-wide web. *Computer Networks and ISDN Systems*, 25(4-5):454 – 459.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022.

Bollen, D., Knijnenburg, B. P., Willemsen, M. C., and Graus, M. (2010). Understanding choice overload in recommender systems. In *Proceedings of the 4th ACM conference on Recommender systems*, RecSys '10, pages 63–70, New York, NY, USA. ACM.

Breese, J. S., Heckerman, D., and Kadie, C. (1998). Empirical analysis of predictive algorithms for collaborative filtering. In *Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence*, UAI'98, pages 43–52, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Brin, S. and Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. *Computer Networks*, 30(1-7):107–117.

Burke, R. (2002). Hybrid recommender systems: Survey and experiments. *User Modeling and User-Adapted Interaction*, 12(4):331–370.

Campos, P., Díez, F., and Cantador, I. (2014). Time-aware recommender systems: a comprehensive survey and analysis of existing evaluation protocols. *User Modeling and User-Adapted Interaction*, 24(1-2):67–119.

Cantador, I., Bellogín, A., and Vallet, D. (2010). Content-based recommendation in social tagging systems. In *Proceedings of the 4th ACM Conference on Recommender Systems*, RecSys '10, pages 237–240, New York, NY, USA. ACM.

Cantador, I., Bellogín, A., and Castells, P. (2008). News@hand: A semantic web approach to recommending news. In Nejdl, W., Kay, J., Pu, P., and Herder, E., editors, *Adaptive Hypermedia and Adaptive Web-Based Systems*, volume 5149 of *Lecture Notes in Computer Science*, pages 279–283. Springer Berlin Heidelberg.

Carbonell, J. and Goldstein, J. (1998). The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '98, pages 335–336, New York, NY, USA. ACM.

Carterette, B. (2011). System effectiveness, user models, and user utility: A conceptual framework for investigation. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '11, pages 903–912, New York, NY, USA. ACM.

Castells, P., Hurley, N., and Vargas, S. (in press). Novelty and diversity in recommender systems. In Ricci, F., Rokach, L., and Shapira, B., editors, *Recommender Systems Handbook 2nd Edition*. Springer US.

Castells, P., Vargas, S., and Wang, J. (2011). Novelty and diversity metrics for recommender systems: Choice, discovery and relevance. In *International Workshop on Diversity in Document Retrieval at the 33rd European Conference on Information Retrieval*, DDR'11.

Celma, O. and Herrera, P. (2008). A new approach to evaluating novel recommendations. In *Proceedings of the 2nd ACM Conference on Recommender Systems*, RecSys '08, pages 179–186, New York, NY, USA. ACM.

Chapelle, O., Ji, S., Liao, C., Velipasaoglu, E., Lai, L., and Wu, S.-L. (2011). Intent-based diversification of web search results: Metrics and algorithms. *Information Retrieval*, 14(6):572–592.

Chapelle, O., Metlzer, D., Zhang, Y., and Grinspan, P. (2009). Expected reciprocal rank for graded relevance. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, CIKM '09, pages 621–630, New York, NY, USA. ACM.

Chen, H. and Karger, D. R. (2006). Less is more: Probabilistic models for retrieving fewer relevant documents. In *Proceedings of the 29th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '06, pages 429–436, New York, NY, USA. ACM.

Chen, L., Wu, W., and He, L. (2013). How personality influences users' needs for recommendation diversity? In *CHI '13 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '13, pages 829–834, New York, NY, USA. ACM.

Clarke, C. L., Craswell, N., Soboroff, I., and Ashkan, A. (2011a). A comparative analysis of cascade measures for novelty and diversity. In *Proceedings of the 4th ACM International Conference on Web Search and Data Mining*, WSDM '11, pages 75–84, New York, NY, USA. ACM.

Clarke, C. L., Kolla, M., Cormack, G. V., Vechtomova, O., Ashkan, A., Büttcher, S., and MacKinnon, I. (2008). Novelty and diversity in information retrieval evaluation. In *Proceedings of the 31st International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '08, pages 659–666, New York, NY, USA. ACM.

Clarke, C. L. A., Craswell, N., and Soboroff, I. (2009). Overview of the TREC 2009 web track. In *Proceedings of the 18th Text REtrieval Conference*, TREC'09.

Clarke, C. L. A., Craswell, N., and Soboroff, I. (2010). Overview of the TREC 2010 web track. In *Proceedings of the 19th Text REtrieval Conference*, TREC'10.

Clarke, C. L. A., Craswell, N., Soboroff, I., and Voorhees, E. M. (2011b). Overview of the TREC 2011 web track. In *Proceedings of the 20th Text REtrieval Conference*, TREC'11.

Clarke, C. L. A., Craswell, N., and Voorhees, E. M. (2012). Overview of the TREC 2012 web track. In *Proceedings of the 21st Text REtrieval Conference*, TREC'12.

Claypool, M., Gokhale, A., Miranda, T., Murnikov, P., Netes, D., and Sartin, M. (1999). Combining content-based and collaborative filters in an online newspaper. In *Proceedings of the ACM SIGIR Workshop Recommender Systems: Algorithms and Evaluation*.

Coelho, F., Devezas, J., and Ribeiro, C. (2013). Large-scale crossmedia retrieval for playlist generation and song discovery. In *Proceedings of the 10th Conference on Open Research Areas in Information Retrieval*, OAIR '13, pages 61–64, Paris, France. CID.

Cremonesi, P., Garzotto, F., and Quadrana, M. (2013). Evaluating top-n recommendations "when the best are gone". In *Proceedings of the 7th ACM Conference on Recommender Systems*, RecSys '13, pages 339–342, New York, NY, USA. ACM.

Cremonesi, P., Koren, Y., and Turrin, R. (2010). Performance of recommender algorithms on top-n recommendation tasks. In *Proceedings of the 4th ACM Conference on Recommender Systems*, RecSys '10, pages 39–46, New York, NY, USA. ACM.

Dang, V. and Croft, W. B. (2012). Diversity by proportionality: An election-based approach to search result diversification. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '12, pages 65–74, New York, NY, USA. ACM.

Desrosiers, C. and Karypis, G. (2011). A comprehensive survey of neighborhood-based recommendation methods. In Ricci, F., Rokach, L., Shapira, B., and Kantor, P. B., editors, *Recommender Systems Handbook*, pages 107–144. Springer US.

Di Noia, T., Ostuni, V. C., Rosati, J., Tomeo, P., and Di Sciascio, E. (2014). An analysis of users' propensity toward diversity in recommendations. In *Proceedings of the 8th ACM Conference on Recommender Systems*, RecSys '14, pages 285–288, New York, NY, USA. ACM.

Ekstrand, M. D., Harper, F. M., Willemsen, M. C., and Konstan, J. A. (2014). User perception of differences in recommender algorithms. In *Proceedings of the 8th ACM Conference on Recommender Systems*, RecSys '14, pages 161–168, New York, NY, USA. ACM.

Ekstrand, M. D., Ludwig, M., Kolb, J., and Riedl, J. T. (2011). Lenskit: A modular recommender framework. In *Proceedings of the 5th ACM Conference on Recommender Systems*, RecSys '11, pages 349–350, New York, NY, USA. ACM.

Elahi, M., Ricci, F., and Rubens, N. (2014). Active learning in collaborative filtering recommender systems. In Hepp, M. and Hoffner, Y., editors, *E-Commerce and Web Technologies*, volume 188 of *Lecture Notes in Business Information Processing*, pages 113–124. Springer International Publishing.

Fleder, D. M. and Hosanagar, K. (2009). Blockbuster culture's next rise or fall: The impact of recommender systems on sales diversity. *Management Science*, 55(5):697–712.

Gantner, Z., Rendle, S., Freudenthaler, C., and Schmidt-Thieme, L. (2011). Mymedialite: A free recommender system library. In *Proceedings of the 5th ACM Conference on Recommender Systems*, RecSys '11, pages 305–308, New York, NY, USA. ACM.

Garcin, F., Faltings, B., Donatsch, O., Alazzawi, A., Bruttin, C., and Huber, A. (2014). Offline and online evaluation of news recommender systems at swissinfo.ch. In *Proceedings of the 8th ACM Conference on Recommender Systems*, RecSys '14, pages 169–176, New York, NY, USA. ACM.

Ge, M., Delgado-Battenfeld, C., and Jannach, D. (2010). Beyond accuracy: Evaluating recommender systems by coverage and serendipity. In *Proceedings of the 4th ACM Conference on Recommender Systems*, RecSys '10, pages 257–260, New York, NY, USA. ACM.

Golbus, P., Pavlu, V., and Aslam, J. (2012). What we talk about when we talk about diversity. In *Proceedings of the International Workshop on Diversity in Document Retrieval at the 5th ACM International Conference on Web Search and Data Mining*, DDR'12, Seattle, Washington, USA.

Goldberg, K., Roeder, T., Gupta, D., and Perkins, C. (2001). Eigentaste: A constant time collaborative filtering algorithm. *Information Retrieval*, 4(2):133–151.

Gunawardana, A. and Shani, G. (2009). A survey of accuracy evaluation metrics of recommendation tasks. *Journal of Machine Learning Research*, 10:2935–2962.

Herlocker, J. L., Konstan, J. A., Borchers, A., and Riedl, J. (1999). An algorithmic framework for performing collaborative filtering. In *Proceedings of the 22nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR 1999, pages 230–237, New York, NY, USA. ACM.

Herlocker, J. L., Konstan, J. A., and Riedl, J. (2000). Explaining collaborative filtering recommendations. In *Proceedings of the 2000 ACM conference on Computer Supported Cooperative Work*, CSCW '00, pages 241–250, New York, NY, USA. ACM.

Herlocker, J. L., Konstan, J. A., Terveen, L. G., and Riedl, J. T. (2004). Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems*, 22(1):5–53.

Hofmann, K., Schuth, A., Bellogin, A., and de Rijke, M. (2014). Effects of position bias on click-based recommender evaluation. In *Proceedings of the 36th European Conference on Information Retrieval*, ECIR'14. Springer.

Hofmann, T. (2004). Latent semantic models for collaborative filtering. *ACM Transactions on Information Systems*, 22(1):89–115.

Hu, B., Zhang, Y., Chen, W., Wang, G., and Yang, Q. (2011). Characterizing search intent diversity into click models. In *Proceedings of the 20th International Conference on World Wide Web*, WWW '11, pages 17–26, New York, NY, USA. ACM.

Hu, Y., Koren, Y., and Volinsky, C. (2008). Collaborative filtering for implicit feedback datasets. In *Proceedings of the 8th IEEE International Conference on Data Mining*, ICDM '08, pages 263–272, Washington, DC, USA. IEEE Computer Society.

Huang, Z., Chen, H., and Zeng, D. (2004). Applying associative retrieval techniques to alleviate the sparsity problem in collaborative filtering. *ACM Transactions on Information Systems*, 22(1):116–142.

Hurley, N. and Zhang, M. (2011). Novelty and diversity in top-N recommendation – analysis and evaluation. *ACM Transactions on Internet Technology*, 10(4):14:1–14:30.

Hurley, N. J. (2013). Personalised ranking with diversity. In *Proceedings of the 7th ACM Conference on Recommender Systems*, RecSys '13, pages 379–382, New York, NY, USA. ACM.

Jannach, D., Lerche, L., and Gda (2013). Re-ranking recommendations based on predicted short-term interests - a protocol and first experiment. In *Proceedings of the AAAI 2013 Workshop on Intelligent Techniques For Web Personalization and Recommender Systems*, ITWP'13. AAAI.

Järvelin, K. and Kekäläinen, J. (2000). Ir evaluation methods for retrieving highly relevant documents. In *Proceedings of the 23rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '00, pages 41–48, New York, NY, USA. ACM.

Joachims, T. (2006). Training linear svms in linear time. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '06, pages 217–226, New York, NY, USA. ACM.

Kabutoya, Y., Iwata, T., Toda, H., and Kitagawa, H. (2013). A probabilistic model for diversifying recommendation lists. In Ishikawa, Y., Li, J., Wang, W., Zhang, R., and Zhang, W., editors, *Web Technologies and Applications*, volume 7808 of *Lecture Notes in Computer Science*, pages 348–359. Springer Berlin Heidelberg.

Karatzoglou, A., Baltrunas, L., Church, K., and Böhmer, M. (2012). Climbing the app wall: Enabling mobile app discovery through context-aware recommendations. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, CIKM '12, pages 2527–2530, New York, NY, USA. ACM.

Karatzoglou, A., Baltrunas, L., and Shi, Y. (2013). Learning to rank for recommender systems. In *Proceedings of the 7th ACM Conference on Recommender Systems*, RecSys '13, pages 493–494, New York, NY, USA. ACM.

Kluver, D. and Konstan, J. A. (2014). Evaluating recommender behavior for new users. In *Proceedings of the 8th ACM Conference on Recommender Systems*, RecSys '14, pages 121–128, New York, NY, USA. ACM.

Knijnenburg, B., Willemsen, M., Gantner, Z., Soncu, H., and Newell, C. (2012). Explaining the user experience of recommender systems. *User Modeling and User-Adapted Interaction*, 22(4-5):441–504.

Kohavi, R., Longbotham, R., Sommerfield, D., and Henne, R. (2009). Controlled experiments on the web: survey and practical guide. *Data Mining and Knowledge Discovery*, 18(1):140–181.

Koren, Y., Bell, R., and Volinsky, C. (2009). Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37.

Küçüktunç, O., Saule, E., Kaya, K., and Çatalyürek, U. V. (2013). Diversified recommendation on graphs: Pitfalls, measures, and algorithms. In *Proceedings of the 22nd International Conference on World Wide Web*, WWW '13, pages 715–726, Republic and Canton of Geneva, Switzerland. International World Wide Web Conferences Steering Committee.

Lang, K. (1995). Newsweeder: Learning to filter Netnews. In *Proceedings of the 12th International Conference on Machine Learning*, ICML'95, pages 331–339, New York, NY, USA. ACM.

Lathia, N., Hailes, S., Capra, L., and Amatriain, X. (2010). Temporal diversity in recommender systems. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '10, pages 210–217, New York, NY, USA. ACM.

Lee, K. and Lee, K. (2013). Using experts among users for novel movie recommendations. *Journal of Computing Science and Engineering*, 7(1):21–29.

Lemire, D. and Maclachlan, A. (2005). Slope one predictors for online rating-based collaborative filtering. In *Proceedings of SIAM Data Mining*, SDM'05.

Levinson, D. (1998). *Ethnic Groups Worldwide: A ready Reference Handbook*. Oryx Press.

Linden, G., Smith, B., and York, J. (2003). Amazon.com recommendations: Item-to-item collaborative filtering. *IEEE Internet Computing*, 7(1):76–80.

Liu, T. (2009). Learning to rank for information retrieval. *Foundations and Trends in Information Retrieval*, 3(3):225–331.

Lops, P., de Gemmis, M., and Semeraro, G. (2011). Content-based recommender systems: State of the art and trends. In Ricci, F., Rokach, L., Shapira, B., and Kantor, P. B., editors, *Recommender Systems Handbook*, pages 73–105. Springer US.

Lubatkin, M. and Chatterjee, S. (1994). Extending modern portfolio theory into the domain of corporate diversification: Does it apply? *The Academy of Management Journal*, 37(1):109–136.

Malinowski, J., Keim, T., Wendt, O., and Weitzel, T. (2006). Matching people and jobs: A bilateral recommendation approach. In *Proceedings of the 39th Hawaii International Conference on System Sciences*, volume 6 of *HICSS '06*, pages 137c–137c.

Marlin, B. M. and Zemel, R. S. (2009). Collaborative prediction and ranking with non-random missing data. In *Proceedings of the 3rd ACM Conference on Recommender Systems*, RecSys '09, pages 5–12, New York, NY, USA. ACM.

McAlister, L. and Pessemier, E. A. (1982). Variety seeking behaviour: and interdisciplinary review. *Journal of Consumer Research*, 9(3):311–322.

McFee, B., Bertin-Mahieux, T., Ellis, D. P., and Lanckriet, G. R. (2012). The million song dataset challenge. In *Proceedings of the 21st International Conference Companion on World Wide Web*, WWW '12 Companion, pages 909–916, New York, NY, USA. ACM.

McNee, S. M., Riedl, J., and Konstan, J. A. (2006). Being accurate is not enough: how accuracy metrics have hurt recommender systems. In *CHI '06 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '06, pages 1097–1101, New York, NY, USA. ACM.

Mei, Q., Guo, J., and Radev, D. (2010). Divrank: The interplay of prestige and diversity in information networks. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '10, pages 1009–1018, New York, NY, USA. ACM.

Moffat, A. and Zobel, J. (2008). Rank-biased precision for measurement of retrieval effectiveness. *ACM Transactions on Information Systems*, 27(1):2:1–2:27.

Mooney, R. J. and Roy, L. (2000). Content-based book recommending using learning for text categorization. In *Proceedings of the 5th ACM Conference on Digital Libraries*, DL '00, pages 195–204, New York, NY, USA. ACM.

Murakami, T., Mori, K., and Orihara, R. (2008). Metrics for evaluating the serendipity of recommendation lists. In Satoh, K., Inokuchi, A., Nagao, K., and Kawamura, T., editors, *New Frontiers in Artificial Intelligence*, volume 4914 of *Lecture Notes in Computer Science*, pages 40–46. Springer Berlin Heidelberg.

Niemann, K. and Wolpers, M. (2013). A new collaborative filtering approach for increasing the aggregate diversity of recommender systems. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '13, pages 955–963, New York, NY, USA. ACM.

Onuma, K., Tong, H., and Faloutsos, C. (2009). Tangent: A novel, 'surprise me', recommendation algorithm. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '09, pages 657–666, New York, NY, USA. ACM.

Ostuni, V. C., Di Noia, T., Di Sciascio, E., and Mirizzi, R. (2013). Top-n recommendations from implicit feedback leveraging linked open data. In *Proceedings of the 7th ACM Conference on Recommender Systems*, RecSys '13, pages 85–92, New York, NY, USA. ACM.

Owen, S., Anil, R., Dunning, T., and Friedman, E. (2011). *Mahout in Action*. Manning Publications Co.

Pariser, E. (2011). *The Filter Bubble: What the Internet Is Hiding from You*. Penguin Group , The.

Park, Y.-J. and Tuzhilin, A. (2008). The long tail of recommender systems and how to leverage it. In *Proceedings of the 2th ACM Conference on Recommender Systems*, RecSys 2008, pages 11–18, New York, NY, USA. ACM.

Patil, G. P. and Taillie, C. (1982). Diversity as a concept and its measurement. *Journal of the American Statistical Association*, 77(379):548–561.

Pazzani, M. (1999). A framework for collaborative, content-based and demographic filtering. *Artificial Intelligence Review*, 13(5-6):393–408.

Pazzani, M. and Billsus, D. (1997). Learning and revising user profiles: The identification of interesting web sites. *Machine Learning*, 27(3):313–331.

Pazzani, M. and Billsus, D. (2007). Content-based recommendation systems. In Brusilovsky, P., Kobsa, A., and Nejdl, W., editors, *The Adaptive Web*, volume 4321 of *Lecture Notes in Computer Science*, pages 325–341. Springer Berlin Heidelberg.

Pizzato, L., Rej, T., Chung, T., Koprinska, I., and Kay, J. (2010). Recon: A reciprocal recommender for online dating. In *Proceedings of the 4th ACM Conference on Recommender Systems*, RecSys '10, pages 207–214, New York, NY, USA. ACM.

Popescul, A., Ungar, L. H., Pennock, D. M., and Lawrence, S. (2001). Probabilistic models for unified collaborative and content-based recommendation in sparse-data environments. In *Proceedings of the 17th Conference in Uncertainty in Artificial Intelligence*, UAI'01, pages 437–444.

Pu, P., Chen, L., and Hu, R. (2011). A user-centric evaluation framework for recommender systems. In *Proceedings of the 5th ACM Conference on Recommender Systems*, RecSys '11, pages 157–164, New York, NY, USA. ACM.

Radlinski, F., Kleinberg, R., and Joachims, T. (2008). Learning diverse rankings with multi-armed bandits. In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, pages 784–791, New York, NY, USA. ACM.

Rendle, S. and Freudenthaler, C. (2014). Improving pairwise learning for item recommendation from implicit feedback. In *Proceedings of the 7th ACM International Conference on Web Search and Data Mining*, WSDM '14, pages 273–282, New York, NY, USA. ACM.

Resnick, P., Iacovou, N., Suchak, M., Bergstrom, P., and Riedl, J. (1994). Grouplens: An open architecture for collaborative filtering of netnews. In *Proceedings of the 1994 ACM Conference on Computer Supported Cooperative Work*, CSCW '94, pages 175–186, New York, NY, USA. ACM.

Resnick, P. and Varian, H. R. (1997). Recommender systems. *Communications of the ACM*, 40(3):56–58.

Ribeiro, M. T., Lacerda, A., Veloso, A., and Ziviani, N. (2012). Pareto-efficient hybridization for multi-objective recommender systems. In *Proceedings of the 6th ACM Conference on Recommender Systems*, RecSys '12, pages 19–26, New York, NY, USA. ACM.

Ricci, F., Rokach, L., and Shapira, B. (2011). Introduction to recommender systems handbook. In Ricci, F., Rokach, L., Shapira, B., and Kantor, P. B., editors, *Recommender Systems Handbook*, pages 1–35. Springer US.

Robertson, S. E. (1997). The probability ranking principle in ir. In Sparck Jones, K. and Willett, P., editors, *Readings in Information Retrieval*, pages 281–286. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Said, A. and Bellogín, A. (2014). Rival: A toolkit to foster reproducibility in recommender system evaluation. In *Proceedings of the 8th ACM Conference on Recommender Systems*, RecSys '14, pages 371–372, New York, NY, USA. ACM.

Said, A., Fields, B., Jain, B. J., and Albayrak, S. (2013). User-centric evaluation of a k-furthest neighbor collaborative filtering recommender algorithm. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work*, CSCW '13, pages 1399–1408, New York, NY, USA. ACM.

Said, A., Larson, M., Tikk, D., Cremonesi, P., Karatzoglou, A., Hopfgartner, F., Turrin, R., and Geurts, J. (2014). User-item reciprocity in recommender systems: Incentivizing the crowd. In *Proceedings of the 22nd Conference on User Modeling, Adaptation, and Personalization*, UMAP'14.

Salakhutdinov, R., Mnih, A., and Hinton, G. (2007). Restricted boltzmann machines for collaborative filtering. In *Proceedings of the 24th International Conference on Machine Learning*, ICML '07, pages 791–798, New York, NY, USA. ACM.

Santos, R. L., Macdonald, C., and Ounis, I. (2010a). Exploiting query reformulations for web search result diversification. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pages 881–890, New York, NY, USA. ACM.

Santos, R. L., Macdonald, C., and Ounis, I. (2010b). Selectively diversifying web search results. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, CIKM '10, pages 1179–1188, New York, NY, USA. ACM.

Santos, R. L., Macdonald, C., and Ounis, I. (2012). On the role of novelty for search result diversification. *Information Retrieval*, 15(5):478–502.

Sarwar, B., Karypis, G., Konstan, J., and Riedl, J. (2001). Item-based collaborative filtering recommendation algorithms. In *Proceedings of the 10th International Conference on World Wide Web*, WWW '01, pages 285–295, New York, NY, USA. ACM.

Schein, A. I., Popescul, A., Ungar, L. H., and Pennock, D. M. (2002). Methods and metrics for cold-start recommendations. In *Proceedings of the 25th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '02, pages 253–260, New York, NY, USA. ACM.

Shani, G. and Gunawardana, A. (2011). Evaluating recommendation systems. In Ricci, F., Rokach, L., Shapira, B., and Kantor, P. B., editors, *Recommender Systems Handbook*, pages 257–297. Springer US.

Shani, G., Heckerman, D., and Brafman, R. I. (2005). An mdp-based recommender system. *Journal of Machine Learning Research*, 6:1265–1295.

Shannon, C. (1948). A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423 & 623–656.

Shardanand, U. and Maes, P. (1995). Social information filtering: Algorithms for automating &ldquo;word of mouth&rdquo;. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '95, pages 210–217, New York, NY, USA. ACM Press/Addison-Wesley Publishing Co.

Shi, Y., Karatzoglou, A., Baltrunas, L., Larson, M., Oliver, N., and Hanjalic, A. (2012a). Climf: Learning to maximize reciprocal rank with collaborative less-is-more filtering. In *Proceedings of the 6th ACM Conference on Recommender Systems*, RecSys '12, pages 139–146, New York, NY, USA. ACM.

Shi, Y., Larson, M., and Hanjalic, A. (2010). List-wise learning to rank with matrix factorization for collaborative filtering. In *Proceedings of the 4th ACM Conference on Recommender Systems*, RecSys '10, pages 269–272, New York, NY, USA. ACM.

Shi, Y., Larson, M., and Hanjalic, A. (2013). Unifying rating-oriented and ranking-oriented collaborative filtering for improved recommendation. *Information Sciences*, 229:29–39.

Shi, Y., Zhao, X., Wang, J., Larson, M., and Hanjalic, A. (2012b). 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*, SIGIR '12, pages 175–184, New York, NY, USA. ACM.

Shih, W., Kaufman, S., and Spinola, D. (2007). Netflix. *Harvard Business School Case*, 9:607-138.

Slivkins, A., Radlinski, F., and Gollapudi, S. (2010). Learning optimally diverse rankings over large document collections. In Fürnkranz, J. and Joachims, T., editors, *Proceedings of the 27th International Conference on Machine Learning (ICML-10), June 21-24, 2010, Haifa, Israel*, pages 983–990. Omnipress.

Smyth, B. and McClave, P. (2001). Similarity vs. diversity. In *Proceedings of the 4th International Conference on Case-Based Reasoning*, ICCBR'01, pages 347–361, London, UK. Springer-Verlag.

Soboroff, I. and Nicholas, C. (1999). Combining content and collaboration in text filtering. In *Proceedings of the IJCAI'99 Workshop on Machine Learning for Information Filtering*.

Steck, H. (2011). Item popularity and recommendation accuracy. In *Proceedings of the 5th ACM Conference on Recommender Systems*, RecSys '11, pages 125–132, New York, NY, USA. ACM.

Steck, H. (2013). Evaluation of recommendations: Rating-prediction and ranking. In *Proceedings of the 7th ACM Conference on Recommender Systems*, RecSys '13, pages 213–220, New York, NY, USA. ACM.

Su, R., Yin, L., Chen, K., and Yu, Y. (2013). Set-oriented personalized ranking for diversified top-n recommendation. In *Proceedings of the 7th ACM Conference on Recommender Systems*, RecSys '13, pages 415–418, New York, NY, USA. ACM.

Sweeney, S., Crestani, F., and Losada, D. E. (2008). 'show me more': Incremental length summarisation using novelty detection. *Information Processing & Management*, 44(2):663–686.

Szlávik, Z., Kowalczyk, W., and Schut, M. (2011). Diversity measurement of recommender systems under different user choice models. In *Proceedings of the 5th AAAI Conference on Weblogs and Social Media*, ICWSM 2011. The AAAI Press.

Takács, G. and Tikk, D. (2012). Alternating least squares for personalized ranking. In *Proceedings of the 6th ACM Conference on Recommender Systems*, RecSys '12, pages 83–90, New York, NY, USA. ACM.

Toffler, A. (1970). *Future shock*. Random House.

Ungar, L. and Foster, D. (1998). Clustering methods for collaborative filtering. Technical report, AAAI.

Vargas, S. (2011). New approaches to diversity and novelty in recommender systems. In *Proceedings of the 4th BCS-IRSG Conference on Future Directions in Information Access*, FDIA'11, pages 8–13, Swinton, UK. British Computer Society.

Vargas, S. (2014). Novelty and diversity enhancement and evaluation in recommender systems and information retrieval. In *Proceedings of the 37th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '14, pages 1281–1281, New York, NY, USA. ACM.

Vargas, S., Baltrunas, L., Karatzoglou, A., and Castells, P. (2014). Coverage, redundancy and size-awareness in genre diversity for recommender systems. In *Proceedings of the 8th ACM Conference on Recommender Systems*, RecSys '14, pages 209–216, New York, NY, USA. ACM.

Vargas, S. and Castells, P. (2011). Rank and relevance in novelty and diversity metrics for recommender systems. In *Proceedings of the 5th ACM Conference on Recommender Systems*, RecSys '11, pages 109–116, New York, NY, USA. ACM.

Vargas, S. and Castells, P. (2012). Diversificación en sistemas de recomendación a partir de sub-perfiles de usuario. In *II Congreso Español de Recuperación de Información*, CERI'12.

Vargas, S. and Castells, P. (2013). Exploiting the diversity of user preferences for recommendation. In *Proceedings of the 10th Conference on Open Research Areas in Information Retrieval*, OAIR '13, pages 129–136, Paris, France. CID.

Vargas, S. and Castells, P. (2014a). Improving sales diversity by recommending users to items. In *Proceedings of the 8th ACM Conference on Recommender Systems*, RecSys '14, pages 145–152, New York, NY, USA. ACM.

Vargas, S. and Castells, P. (2014b). Vecindarios inversos para la mejora de la novedad en filtrado colaborativo. In *III Congreso Español de Recuperación de Información*, CERI'14.

Vargas, S., Castells, P., and Vallet, D. (2011). Intent-oriented diversity in recommender systems. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '11, pages 1211–1212, New York, NY, USA. ACM.

Vargas, S., Castells, P., and Vallet, D. (2012a). Explicit relevance models in intent-oriented information retrieval diversification. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '12, pages 75–84, New York, NY, USA. ACM.

Vargas, S., Castells, P., and Vallet, D. (2012b). On the suitability of intent spaces for IR diversification. In *Proceedings of the International Workshop on Diversity in Document Retrieval at the 5th ACM International Conference on Web Search and Data Mining*, DDR'12, Seattle, Washington, USA.

Vargas, S., Santos, R. L. T., Macdonald, C., and Ounis, I. (2013). Selecting effective expansion terms for diversity. In *Proceedings of the 10th Conference on Open Research Areas in Information Retrieval*, OAIR '13, pages 69–76, Paris, France. CID.

Veloso, A., Ribeiro, M., Lacerda, A., Moura, E., Hata, I., and Ziviani, N. (in press). Multi-objective pareto-efficient approaches for recommender systems. *ACM Transactions on Information Systems*.

Verstrepen, K. and Goethals, B. (2014). Unifying nearest neighbors collaborative filtering. In *Proceedings of the 8th ACM Conference on Recommender Systems*, RecSys '14, pages 177–184, New York, NY, USA. ACM.

Wang, J. (2009). Mean-variance analysis: A new document ranking theory in information retrieval. In *Proceedings of the 31th European Conference on IR Research on Advances in Information Retrieval*, ECIR '09, pages 4–16, Berlin, Heidelberg. Springer-Verlag.

Wang, J., Robertson, S., Vries, A. P., and Reinders, M. J. (2008). Probabilistic relevance ranking for collaborative filtering. *Information Retrieval*, 11(6):477–497.

Wang, J. and Zhu, J. (2009). Portfolio theory of information retrieval. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '09, pages 115–122, New York, NY, USA. ACM.

Weimer, M., Karatzoglou, A., Le, Q. V., and Smola, A. J. (2007). COFI RANK - maximum margin matrix factorization for collaborative ranking. In *Proceedings of the 21st Annual Conference on Neural Information Processing Systems*, NIPS'07.

Welch, M. J., Cho, J., and Olston, C. (2011). Search result diversity for informational queries. In *Proceedings of the 20th International Conference on World Wide Web*, WWW '11, pages 237–246, New York, NY, USA. ACM.

Winoto, P. and Tang, T. Y. (2010). The role of user mood in movie recommendations. *Expert Systems with Applications*, 37(8):6086– 6092.

Wu, H., Cui, X., He, J., Li, B., and Pei, Y. (2014). On improving aggregate recommendation diversity and novelty in folksonomy-based social systems. *Personal and Ubiquitous Computing*, 18(8):1855–1869.

Yu, K., Schwaighofer, A., Tresp, V., Xu, X., and Kriegel, H. (2004). Probabilistic memory-based collaborative filtering. *IEEE Transactions on Knowledge and Data Engineering*, 16(1):56–69.

Yue, Y. and Joachims, T. (2008). Predicting diverse subsets using structural svms. In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, pages 1224–1231, New York, NY, USA. ACM.

Zadeh, R. B. and Carlsson, G. (2013). Dimension independent matrix square using mapreduce. *CoRR*, abs/1304.1467.

Zhai, C. X., Cohen, W. W., and Lafferty, J. (2003). Beyond independent relevance: Methods and evaluation metrics for subtopic retrieval. In *Proceedings of the 26th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '03, pages 10–17, New York, NY, USA. ACM.

Zhang, M. and Hurley, N. (2008). Avoiding monotony: Improving the diversity of recommendation lists. In *Proceedings of the 2nd ACM Conference on Recommender Systems*, RecSys '08, pages 123–130, New York, NY, USA. ACM.

Zhang, M. and Hurley, N. (2009). Novel item recommendation by user profile partitioning. In *Proceedings of the 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology*, WI-IAT '09, pages 508–515, Washington, DC, USA. IEEE Computer Society.

Zhang, Y. C., Séaghdha, D. O., Quercia, D., and Jambor, T. (2012). Auralist: Introducing serendipity into music recommendation. In *Proceedings of the 5th ACM*

International Conference on Web Search and Data Mining, WSDM '12, pages 13–22, New York, NY, USA. ACM.

Zhao, X., Niu, Z., and Chen, W. (2013). Opinion-based collaborative filtering to solve popularity bias in recommender systems. In *Database and Expert Systems Applications*, volume 8056 of *Lecture Notes in Computer Science*, pages 426–433. Springer Berlin Heidelberg.

Zhou, T., Kuscsik, Z., Liu, J.-G., Medo, M., Wakeling, J. R., and Zhang, Y.-C. (2010). Solving the apparent diversity-accuracy dilemma of recommender systems. *Proceedings of the National Academy of Sciences*, 107(10):4511–4515.

Ziegler, C.-N., McNee, S. M., Konstan, J. A., and Lausen, G. (2005). Improving recommendation lists through topic diversification. In *Proceedings of the 14th International Conference on World Wide Web*, WWW '05, pages 22–32, New York, NY, USA. ACM.