Universidad Autónoma de Madrid
Escuela Politécnica Superior
Departamento de Ingeniería Informática

Novelty and Diversity Enhancement and Evaluation in Recommender Systems

Saúl Vargas Sandoval
Supervisor: Pablo Castells Azpilicueta

Trabajo Fin de Máster

Programa Oficial de Posgrado en Ingeniería Informática y de Telecomunicación.
Universidad Autónoma de Madrid
Abril 2012

Ninguna cosa despierta tanto el bullicio del pueblo como la novedad.

Francisco de Quevedo

La unidad es la variedad, y la variedad en la unidad es la ley suprema del Universo

Isaac Newton

Abstract

Novelty and diversity as relevant dimensions of retrieval quality are receiving increasing attention in the Information Retrieval and Recommender Systems fields. Both problems have nonetheless been approached under different views and formulations in Information Retrieval and Recommender Systems respectively, giving rise to different models, methodologies, and metrics, with little convergence between both fields. We find considerable room for research towards the formalization of diversification methods, evaluation methodologies, and metrics. Furthermore, we ask ourselves whether there should be some natural connection between the perspectives on diversity in Information Retrieval and Recommender Systems, given that recommendation is after all an information retrieval problem.

In the present work we propose an Information Retrieval approach to the evaluation and enhacement of novelty and diversity in Recommender Systems. We draw models and solutions from text retrieval and apply them to recommendation tasks in such a way that the recent advances achieved in the former can be leveraged for the latter.

We also propose a new formalization and unification of the way novelty and diversity are evaluated on Recommender Systems, considering rank and relevance as additional and meaningful aspects for the evaluation of recommendation lists. We propose a framework that includes and unifies the main state of the art metrics for novelty and diversity in Recommender Systems, generalizing and extending them with further properties and flexibility in configuration.

Our contributions are tested with standard Recommender Systems collections, in order to validate our proposals and provide further insights.

Contents

1. Introduction

1.1 Motivation

Until recently, research in Information Retrieval (IR) and Recommender Systems (RS) has focused almost exclusively on achieving accuracy, i.e., retrieving the most individually relevant documents or items for the needs of a query or a user, respectively. However, novelty and diversity as relevant dimensions of retrieval quality are receiving increasing attention in both IR and RS. The problem has nonetheless been approached under different views and formulations in both fields, giving rise to different models, methodologies, and metrics, with little convergence between both fields.

Recommender Systems can be seen as a particular case of personalized Information Retrieval where there is no explicit query, but just implicit information about the user's interests. Recommendation tasks generally involve a large set of items –such as books, movies or songs– and a large set of users to which the system provides suggestions of items they may enjoy or benefit from. Recommender systems technologies have experienced a considerable development with significant impact and introduction in commercial applications.

The primary objective of every RS is to satisfy the seller's interests by satisfying the customer's interests. The classical approach for this task has been to predict a score for an item the user has not judged or accessed, and then present these new items in decreasing order of score. Nevertheless, this mechanism alone is usually not enough to actually satisfy the user's interests. For example, if a system recommends items based on their popularity, it is likely not doing a task the user could not have done by herself – even if the user happens to like the items, the chances that she had already heard about them are high, whereby the recommendation is of very marginal –if any– use. As another case, a very accurate system could return a set of monothematic items matching the user known themes or interests. This approximation may also fail since, albeit accurately matching the user's preferences, the whole set of recommended items may be perceived as one –consider the case of a music recommendation algorithm that only returns songs of the same artist. The key in these situations is that novelty and diversity should be also considered in the quality assessment of a RS, as accuracy alone gives a very partial account of the actual system's utility.

The problem of results diversity has been already addressed in IR, but from a different angle. The diversity dimension of search results is being researched in the IR field as a means to address the ambiguity and/or underspecification involved in user queries. Current approaches to enhance and evaluate the diversity of search results use concepts such as query intents and document similarity. Query intents can be seen as the different meanings or purposes an underspecified query can represent. Taxonomies and query logs have been used for discovering and describing these intents. The identification of query intents and interpretations is then used to discover categories or refinements which may suit a query. Maximizing the range of categories covered by returned documents is a means to cope with the initial ambiguity of a query.

In RS the focus lies on broadening the offer of recommended items to present to the user (diversity), and promoting less widely known (so-called long-tail) items (novelty),

or items a specific target user is unfamiliar with (unexpectedness). There has been some research in this area as well, and a raising concern for the importance of novelty and diversity in the RS community. However, we find considerable room for research towards the formalization of diversification methods, evaluation methodologies, and metrics in RS, compared to the level of convergence and standardization that is being achieved in the IR community. Furthermore, we ask ourselves whether there should be some natural connection between the perspectives on diversity in IR and RS, given that recommendation is after all a retrieval problem.

1.2 Problem definition

1.2.1 The Recommendation Task

We provide first a brief overview of the general task of a recommender system, and we introduce some notation that shall be used in the following sections. Given a user $u \in \mathcal{U}$ and a set of items $i \in \mathcal{I}$ the task consists on retrieving items the user may like or benefit from. Although some personal information about the user could be used, in general the predictions are generated from the user profile (which we shall denote as **u**), i.e., those items the user has previously interacted with, showing some evidence of her interest for them.

The interaction between a user and an item may consist of an explicit rating $r(u, i)$ (which may be binary –"liked" or "not liked"– or gradual, e.g. one to five stars), or just of item access frequencies f_{ui}, in which the potential interest for the item is evidenced more implicitly. Most of the RS community has focused on the rating case, specifically in the task of rating prediction (Figure 1), which tries to learn a rating prediction function $\hat{r}: \mathcal{U} \times \mathcal{I} \rightarrow [1, \ldots, r_{max}]$. However, we will consider a more general case of a top-N recommendation, in which the goal of the recommender is to retrieve a list $R \in \mathcal{I}^N$ of useful items for each user. Traditionally the order of the presented items is given by their individual interest (relevance) predictions. This approach needs not be optimal, as we will show in the next section.

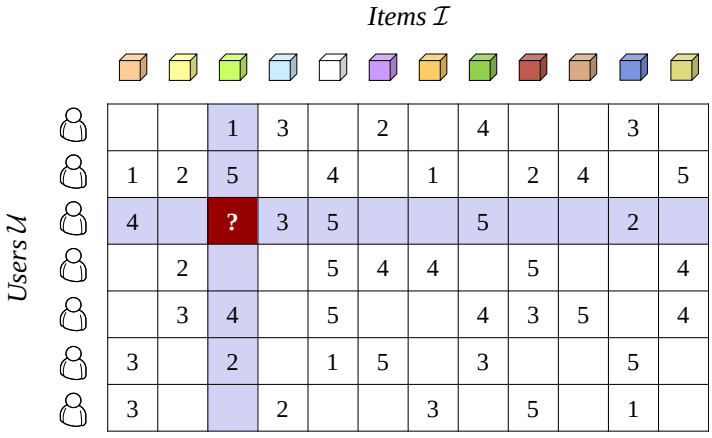

Figure 1. The recommendation problem as a rating prediction task

Based on how the information of the user profile is employed, recommender systems can be divided in three categories:

- Content-based (CB): the recommender will retrieve items whose content is similar to those of the profile.

- Collaborative Filtering (CF): the recommender will retrieve items based on connections or similarities between user profiles.
- Hybrid approaches, combining CB and CF.

The CF approach is specially interesting in scenarios with a large user community with high interaction with items in the collection, and the content of the items is incomplete or hard to work with. Popular CF algorithms include nearest-neighbors, rating matrices factorization (Koren, Bell, & Volinksy, 2009) and probabilistic latent semantic analysis (Hofmann, 2004). A broad and informative survey on the topic can be found in (Adomavicius & Tuzhilin, 2005), which can be complemented with (Koren, Bell, & Volinksy, 2009) for matrix factorization. We provide in the next section a quick overview of the aforementioned methods.

Additionally, we will also consider that items usually have some categorical information associated with them, to which we shall refer as features. Throughout this work we consider a homogeneous set of item features $\mathcal{F}$. In particular, for each item i we denote its subset of features as **i**.

1.2.2 Overview of Some Collaborative Filtering Algorithms

One of the most used solutions to the recommendation tasks is the family of *nearest neighbors* algorithms (kNN), which focuses on the similarity of ratings between users (user-based) or items (item-based) to predict ratings. In its simplest form, user-based approaches consist on predicting ratings for a user u based on the combination of other users' predictions weighted by their similarity with the target user:

$$\hat{r}(u,i) = C \sum_{v \in \mathcal{N}_u} sim(u,v)\, r(v,i)$$

where $\mathcal{N}_u$ is a fixed size neighborhood of most similar users with respect to u and $C = 1/\sum_{v \in \mathcal{N}_u} |sim(u,v)|$ is a normalization constant. The similarity component $sim(u,v)$ is usually computed with the cosine or the correlation between the ratings for common items between users u and v. Item-based alternatives follow a very similar approach, exchanging the role of users and items in rating estimation.

Matrix factorization (MF) (Koren, Bell, & Volinksy, 2009) approaches consider the rating data as part of an incomplete rating matrix and seek to minimize explicitly the mean squared error of rating predictions. Inspired by the SVD factorization of matrices, MF algorithms obtain –in its most basic approach– a decomposition of the rating data in two matrices $P \in \mathbb{R}^{|\mathcal{U}|,k}$ and $Q \in \mathbb{R}^{|\mathcal{I}|,k}$ (each row corresponding to a user or item, respectively) such that

$$P, Q = \underset{P \in \mathbb{R}^{|\mathcal{U}|,k}, Q \in \mathbb{R}^{|\mathcal{I}|,k}}{\arg\min} \sum_{u,i : r(u,i) \neq \emptyset} (r(u,i) - P_u\, Q_i^t)^2$$

where $r(u,i) \neq \emptyset$ denotes a known rating. Conceptually, both matrices capture the latent factors in a k-dimensional latent space –with $k \ll |\mathcal{U}|, |\mathcal{I}|$– that condenses the information about the known rating data and predicts unknown ratings:

$$\hat{r}(u,i) = P_u\, Q_i^t$$

Commonly gradient descent or alternating least squares techniques have been used to estimate the factorization.

An alternative way to extract latent characteristics from user-item interaction data is the *probabilistic latent semantic analysis* (pLSA) (Hofmann, 2004), which seeks to learn a model of latent random variables. For the case of binary implicit information, pLSA considers a set $\{z\}$ of hidden variables and learns a model $\theta = \{p(z|u), p(i|z)\}$ such that it can estimate the probability of an item i being chosen from the user u:

$$p(i|u,\theta) = \sum_{z} p(i|z)p(z|u)$$

The model is learnt using an expectation-maximization algorithm that tries to maximize the log-likelihood of the known data:

$$\theta = \arg\min_{\theta} \sum_{u,i} \log p(i|u,\theta)$$

1.2.3 Novelty and Diversity in Recommendations

When using a recommender system such as those of online stores (Amazon.com, Netflix, etc.) one might experience the problem depicted in Figure 2. Since the user profile is composed of a couple of Beatles' albums, a recommendation engine focused solely on accuracy may provide a list composed mainly of other albums of the Beatles and a couple of other authors (Pink Floyd, Bob Dylan). Although it is highly probable that the user will also like the recommended albums, it is clear that the recommendation is not very useful in the sense of:

- the lack of diversity, probably a smaller sample of albums from the Beatles would have been as useful to discover the work of the band and would have given space for other interesting music from other authors; and
- the lack of novelty, since the Beatles are a massive word-wide known band for which a recommender system is not even required.

Figure 2. A not so useful recommendation

This situation opens two immediate questions: ¿why is this happening? ¿how to solve it? Typically, recommender systems have been trained towards minimizing the prediction error, so aspects like redundancy and obviousness have not been generally considered. Another problem lies in the underspecification of the user profile. Since it only contains album of a single author, a pure CF approach is likely to find most of the connections to other users that will have more albums from the same author. Finally, even when the user had bought or browsed other authors' albums, those of the Beatles are so popular that may be inevitably promoted by a standard recommendation algorithm. Ways to solve this problem are presented in the state of the art (section 2.3) and in our contributions (chapters 3 and 4).

Apart from the decision of which items to present, an additional aspect to consider is the order those are presented. Even when a recommendation consists in a selection of the top-N items with highest-predicted relevance for the target user, the order in which these are presented influences the user perception. Consider the case of Figure 3. Although the top-9 recommended items in lists R_1 and R_2 are the same, the order is different. A user that only considers the first four results of each will find that R_1 is more diverse than R_2.

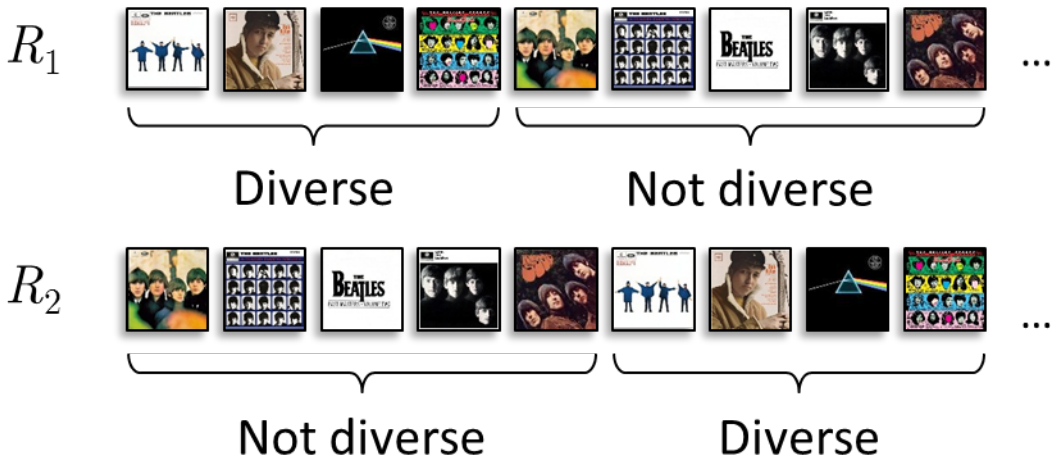

Figure 3. Different recommendation lists with the same items

Another situation to take into account is the case of a system with a high ratio of novel, different items that do not match at all the likes of the user. That would be the case of a random recommender. Although we have seen that a purely accuracy-based recommender may fail in providing a useful recommendation list, it is obvious that accuracy should be kept while enhancing other user-centric dimensions. Consider the hypothetical case depicted in Figure 4. For recommendation lists are compared in terms of accuracy and diversity. While list D is clearly better than B and C, the rest of the combinations are not easily comparable between them in an objective way, and would be dependent on how much each utility dimension (accuracy, diversity, novelty) is valued by the user. In this context, a single metric encompassing both accuracy and novelty or diversity should be useful for a more comprehensive and conclusive comparison of recommendation lists.

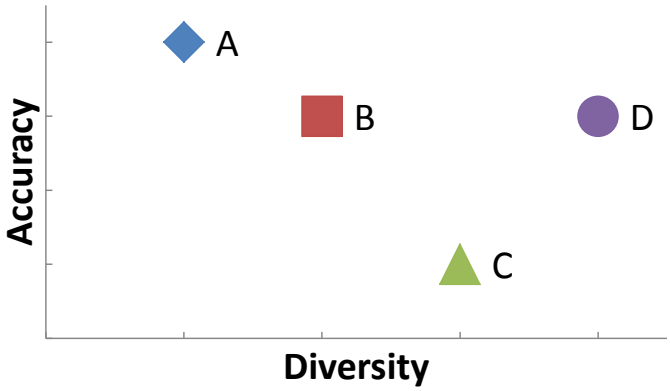

Figure 4. Accuracy-diversity plot for four different recommendations

1.3 Research Goals

The research goals of this work are twofold. First, we aim to identify and analyze the current proposals and trends in the area of interest and others closely related with it, specifically:

- Novelty and Diversity in IR: we study the different definitions and situations where novel and diverse results are desired. Then we analyze specific approaches both to assess and generate novelty and diversity in search results.
- Novelty and Diversity in RS: analogously to IR, we study the contexts and situations that motivate novel and diverse recommendations, analyzing individual contributions to the field, both for evaluation and for improvement.
- IR Metric formalization schemes: we aim to study the latest advances in the formalization of metrics based on utility models derived from how the user examines and consumes items in a result list.

Second, starting from the work done on novelty and diversity for IR and RS, our research seeks progress towards a unification of views, and the identification of essential elements and principles on which a theory of diversity could be built. Concretely, we seek to present two original contributions to the main field of study:

- Adaptation of IR diversity models, algorithms and metrics to RS, by bridging the principles proposed in search diversity and the elements involved in a recommendation scenario.
- Development of a framework for the definition of RS novelty an diversity metrics that unifies different perspectives and state of the art metrics, and supports configurations that take into account the ranking and relevance of recommended items, two aspects not considered by the recommendation diversity metrics reported in the literature.

1.4 Publications

The contributions of this work have had been reported in the following publications, presented in chronological order:

- In (Castells, Vargas, & Wang, 2011) we report a preliminary work on the formalization of metrics for RS presented in chapter 4.

- In (Vargas, Castells, & Vallet, 2011) we explore the adaptation of diversity metrics, techniques, and principles from ad-hoc IR to the recommendation task. The contents of this paper are expanded in chapter 3.
- We present in (Vargas & Castells, 2011) a formal framework for the definition of novelty and diversity metrics that unifies and generalizes several state of the art metrics, relating to the work presented in chapter 4.

As part of the ongoing and future work described in section 5.3, we make a first approach to the problem of the suitability of intent spaces for IR and RS diversification in (Vargas, Castells, & Vallet, 2012).

1.5 Document Structure

The rest of the document continues as follows:

- In chapter 2 a comprehensive study of the state of the art is presented. Three general areas have been identified: evaluation metrics for IR (section 2.1), novelty and diversity in IR (section 2.2) and novelty and diversity in RS (2.3).
- Chapter 3 presents our contributions in adapting the state of the art of diversity in IR to RS.
- Chapter 4 defines a framework for defining novelty and diversity metrics for RS that can consider rank and relevance.
- Chapter 5 offers a summary of the work, stresses the presented contributions, discusses some possible corrections or improvements to the contributions and outlines the future work.

2. State of the Art

In this section we present an extensive analysis of the state of the art to identify and analyze current trends in the study of algorithms and evaluation metrics for IR, novelty and diversity in IR and RS. Since the Recommender Systems field has part of its roots in the Information Retrieval field, we envision the investigation of the work on novelty and diversity in the latter area as an important initial step with potential for advancement in the former.

There has been extensive work in the past decade of research in the IR field on how to assess results list quality from the user's point of view, whether from a pure accuracy-based view or a novelty and diversity-based one. These advancements contrast with the development of models, algorithms and theories in RS, where researchers are still attached to simple error-based evaluation metrics, and there is incomplete consensus in the evaluation methodologies for basic recommendation properties such as ranking or relevance. We therefore find a useful reference in revising the evaluation procedures and metrics in IR.

2.1 Evaluation Metrics for Information Retrieval

Along with online evaluation with real users, current IR evaluation practice relies on standard test collections for offline experimentation, such as the TREC datasets (Voorhees & Karman, 2005), each of them consisting of a vast collection of documents, a set of topics (instances of information needs, containing a detailed description and a query that will be the input of the IR system) for the tested systems to search for and a collection of (binary or graded) relevance judgments of documents for each topic. The yearly TREC campaigns provide a rich variety of medium to large-scale test collections, including datasets for such specialized retrieval tasks as cross-language retrieval, blog retrieval, patent search or diversity search, among many others. These collections allow the IR community to have a common resource for comparison and reproducibility of algorithms and methods for text retrieval.

Among the earliest and most traditionally employed metrics to evaluate IR systems' outputs, *precision* and *recall* measure the ratio of returned relevant documents over the number of returned and relevant documents, respectively, considering binary relevance judgments. Since users usually stop browsing search results early in the ranking, it is usual to take a cutoff position in the result list for the computation of these metrics, below which further returned documents are not considered. Precision and recall *at N* are thus defined as:

$$P@N = \frac{\sum_{k=1}^{N} rel_k}{N}$$

$$recall@N = \frac{\sum_{k=1}^{N} rel_k}{R}$$

where rel_k denotes the binary relevance of document at position k and R the total number of relevant documents.

With time, additional metrics have been developed in the field to overcome the limitations of precision and recall, or complement their properties. Some of them, such as *mean average precision* (MAP), *R-precision*, *F-measure* or *mean reciprocal rank* (MRR) are derivations of the aforementioned metrics. Although these metrics have a very simple formulation and are quite simple for interpretation, they may still lack of some connection with how the result list evaluation is conducted by real users. For example, they cannot handle graded relevance values and do not take into account the fact that documents in low positions, though relevant, have an increasingly smaller probability of being examined by the user than those in higher positions. This led to the definition of metrics such as *normalized discounted cumulative gain* (nDCG) (Järvelin & Kekäläinen, 2002), which apply a rank discount to the cumulated relevance that a result list provides the user with.

More recently, researchers realized that many of the metrics which had been used for decades can be connected to formal models describing how users interact with –and draw benefit from– search results. This has led to a unification and formalization of existing metrics, and to the definition of new ones, such as *expected reciprocal rank* (ERR, see Chapelle, Metzler, Zhang, & Grinspan, 2009), *rank-biased precision* (RBP, see Moffat & Zobel, 2008), and others (Clarke C. , Craswell, Soboroff, & Ashkan, 2011). This strand of progress found a strong and bright connection with the research on so-called click models (Hu, Zhang, Chen, & Wang, 2011), leading to a fertile and innovative variety of metrics, theories, connections and insights. Such modern proposals for evaluation metrics for IR tend to focus on the perceived utility for the user, rather than absolute total relevance per se, where two approaches have been identified:

1) The metric models the way the user makes use of the result list –which is commonly referred to as a user model, for short.
2) The metric determines whether a certain objective has been fulfilled (e.g. the family of *k-call* metrics proposed in Chen & Karger, 2006).

We pay close attention to this recent strand of work, as we shall follow a similar methodology in our development of novelty and diversity metrics for recommender systems. For this reason we provide a brief overview of the latest advances and basic principles in this area in the next two subsections.

2.1.1 Metrics Based on User Models

2.1.1.1 Summary

Initially, let us define metrics based on user models that we will consider:

- *Average Precision* (AP) is defined for binary relevance as:

$$AP = \frac{1}{R}\sum_{k=1}^{N} rel_k \frac{1}{k}\sum_{j=1}^{k} rel_j$$

- *Discounted Cumulative Gain* (DCG) for graded relevance is defined as:

$$DCG = \sum_{k=1}^{N} \frac{rel_k}{\log_2 1+k}$$

  though sometimes rel_k is replaced by $2^{rel_k} - 1$ to emphasize the importance of highly relevant documents. DCG is often normalized by dividing by the ideal DCG, in this case it is generally referred as nDCG.
- *Rank-biased Precision* (RBP) is defined as:

$$RBP = (1-p)\sum_{k=1}^{N} p^{k-1}\, rel_k$$

- *Expected Reciprocal Rank* (ERR) is defined for graded relevance as:

$$ERR = \sum_{k=1}^{N} \frac{1}{k} \prod_{i=1}^{k-1} \big(1 - p(rel|i)\big)\, p(rel|k)$$

where the probability of relevance of the document at position k is defined as

$$p(rel|i) = \frac{2^{rel_i} - 1}{2^{rel_{max}}}$$

2.1.1.2 Browsing models

One of the first elements to model the use of a result list is the *browsing model*, i.e., how a user examines the documents in the list. Browsing models are tightly connected with statistical *click models* (see Craswell, Zoeter, Taylor, & Ramsey, 2008 and Hu, Zhang, Chen, & Wang, 2011), used by search engines to determine how to extract relevance information from click logs taking into account factors such as position, intent, etc., that can bias the perception of relevance.

The most obvious bias in a result list is the position of documents. Browsing models that take into account the position in the ranking of documents are called *position models*. These models suppose that the position of a document in a result list determines critically the probability of the document being examined. Specifically, the *examination hypothesis* states the probability of a document being observed depends on its position in the ranking in a monotonically decreasing way:

$$p(seen|k) \geq p(seen|k+1)$$

where $p(seen|k)$ denotes the probability of the document at position k being seen. This could be the case of DCG (Järvelin & Kekäläinen, 2002) and its normalized variant nDCG where there is a logarithmic-like discount $p(seen|k) = 1/\log_2(1+k)$.

A refinement of the previous is the *cascade hypothesis*, which states that the user examines search results from the top downwards, in order and without skipping any document, until she stops browsing at some point. This can be expressed probabilistically as:

$$p(seen|k) = p(seen|k-1)\, p(cont|k-1)$$

where $p(cont|k)$ is the probability of continuing the examination after document at position k. Usually browsing models based on the cascade hypothesis also assume that the document in the first position is always examined (i.e. $p(seen|1) = 1$), so by recursion the formula can be expressed as:

$$p(seen|k) = \prod_{j=1}^{k-1} p(cont|j)$$

The probability of continuing browsing after a certain position can be modeled in different ways. For example, *rank-biased precision* (RBP) models a user that keeps exploring the result lists but, after each step, reflects a constant *impatience* parameter p representing a constant probability to continue browsing (i.e. probability not to stop) at any position:

$$p(cont|k) = p \;\Rightarrow p(seen|k) = p^{k-1}$$

Other models may take into account the relevance at each position to estimate the probability of continuing. In its most general way this can be decomposed as a marginalization with respect to relevance as follows:

$$p(cont|k) = p(cont|k,rel)\, p(rel|k) + p(cont|k,\neg rel)\, \big(1 - p(rel|k)\big)$$

where the probability of relevance $p(rel|j)$ can be estimated from relevance judgments (e.g. simply $p(rel|j) = rel_j$ for binary judgments). $p(cont|k,rel)$ and $p(cont|k,\neg rel)$ can be estimated in different ways using different assumptions. For example, RBP assumes $cont$ is independent from k, that is, $p(cont|k) = p(cont) = p$, as shown above. The *expected reciprocal rank* metric (ERR, see Chapelle, Metzler, Zhang, & Grinspan, 2009) assumes that once the user has found a relevant document the session stops ($p(cont|k,rel) = 0$) and otherwise the user keeps going ($p(cont|k,\neg rel) = 1$), leading to:

$$p(seen|k) = \prod_{j=1}^{k-1} \big(1 - p(rel|j)\big)$$

2.1.1.3 Utility Accumulation Models

Another aspect to determine in a user model is the *utility accumulation model*, which describes how a user accumulates utility from individual relevant documents. Carterette (SIGIR, 2011) proposes a framework that embraces a broad set of metrics and four families of utility accumulation models. The browsing model considered in this work considers, instead of the probability of a document being examined, the probability $p(k|stop)$ of stopping the session at a certain rank k.

The first described model is the *expected utility model*, in which the derived utility is the expected relevance at stopping rank:

$$M_1 : \sum_{k=1}^{N} rel_k\, p(k|stop)$$

Under this framework, RBP can be understood as the expected utility with stopping probability $p(k|stop) = p^{k-1}\,(1-p)$.

The second model is called *expected total utility model*. In this case, the derived utility is not only that of the document at stopping rank k, but the sum of relevance from documents between positions 1 and k:

$$M_2 : \sum_{k=1}^{N} \sum_{j=1}^{k} rel_j\; p(k|stop)$$

The author finds that DCG can be described as an instantiation of this model, where $p(k|stop) = 1/\log_2(1+k) - 1/\log_2(2+k)$.

A third family is that of the metrics based on *expected effort*. Instead of assessing the utility with accumulation relevance grades, it uses an *effort function* at a given rank $f(k)$ to penalize the higher the stopping position, which in this case will be derived from the relevance of the documents from first to stopping position:

$$M_3 : \sum_{k=1}^{N} f(k)\, p(k|stop)$$

An example of this family is ERR, where $f(k) = 1/k$ and

$$p(k|stop) = p(rel|k) \prod_{j=1}^{k-1} \big(1 - p(rel|k)\big)$$

Finally, the fourth family is the one of the *expected average utility*. This model considers the expected effort of further browsing after a relevant document is found:

$$M_4: \sum_{k=1}^{N} rel_k \sum_{j=k}^{N} f(j)\, p(j|stop) = \sum_{k=1}^{N} f(k)\, p(k|stop) \sum_{j=1}^{k} rel_j$$

Under this view, *average precision* (AP, see Robertson, 2008) is an example of this family with $f(k) = 1/k$ and $p(k|stop) = rel_k/R$ where R is the total number of relevant documents in the collection.

2.1.2 Objective-Based Metrics

The other set of metrics of interest for our research do not aim to replicate how the user explores the results list, but focus instead on whether an objective has been accomplished. While these metrics may not have such a strict formal grounding as the ones analyzed in the previous subsection, they have the advantage to be quite clear to understand, and provide a means to define IR systems by optimizing their values.

Chen & Karger (2006) broadly discuss the *Probability Ranking Principle* (PRP), and contend that it is not optimal for IR metrics besides precision or recall. The authors propose a generalization of the PRP towards an *Expected Metric Principle* (EMP) which aims for optimize directly the expected value of the metric of interest. The metrics introduced in the paper are the family of *k-call* metrics. The k-call metric family provides a binary value for a given ranked list of results of length n for a query, returning 1 if at least k ranked documents are relevant and 0 otherwise. In the case of $k = 1$ (*1-call*) the EMP seeks to optimize:

$$P(rel_0 \cup rel_1 \cup \ldots \cup rel_{n-1} \mid d_0, d_1, \ldots, d_{n-1})$$

Since optimizing this probability would pose a NP-hard problem, it is convenient to apply here a greedy approach in which, for each step i from 0 to $n-1$, one should choose d_i so that

$$P(rel_0 \cup rel_1 \cup \ldots \cup rel_i \mid d_0, d_1, \ldots, d_i)$$

is maximized. One can see that maximizing this probability is equivalent to maximizing:

$$P(rel_i \mid \neg rel_0, \neg rel_1, \ldots, \neg rel_{i-1}, d_0, d_1, \ldots, d_i)$$

That is, maximizing the relevance of the i-th document assuming that the previous results were irrelevant. An IR system optimized for the EMP should therefore be able to change its relevance model to adapt at each step the new assumptions of non relevance of previously retrieved documents.

2.2 Novelty and Diversity in Information Retrieval

As previously discussed, IR research and development has been traditionally focused on accuracy and relevance as targets for satisfying the user information need. However, 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 concepts of diversity and novelty are being increasingly recognized as important ingredients of information value in many application domains.

As defined in (Clarke, et al., 2008) *diversity* is a quality of result lists that helps cope with ambiguity or 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 refers clearly 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 football team, etc. In this case these aspects 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 aspects as *subtopics*, since we shall deal with both in the same way –as generally do prior approaches in the state of the art literature. As a strategy to cope with ambiguity and underspecification, several authors have researched approaches that aim to cover as many subtopics as possible (*subtopic retrieval problem* in Zhai, Cohen, & Lafferty, 2003), while still retaining sufficient relevance to satisfy the user need.

Novelty, on the other side, is defined as the quality of a system that avoids redundancy. When an IR system presents the user two documents with the same or very similar content, it is obvious that one those documents adds little marginal utility with respect to the other. This effect is nonetheless not adequately captured by most standard IR metrics. A novel list should be aware of redundancy detection a promote documents that are different between them.

Note that there is another notion of novelty in IR that deals with the so called *sentence novelty*, concerning how to summarize texts without the need to refer to the original source. Since this notion of novelty has no direct application in a general recommendation scenario, we shall leave it outside the scope of our present study. For more details, Sweeney, Crestani, & Losada (2008) provide a comprehensive study about the topic.

Traditionally, IR research has been built upon the *Probability Rank Prinpicle* (*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, 1977). While this principle has been of great utility in the research and development in IR systems for decades, it does not take into account diversity or novelty at all. This issue has been identified by many authors such as Chen & Karger (2006), Clarke, et al. (2008) or Zhai, Cohen & Lafferty (2003). Consider the case of diversity of the *subtopic retrieval problem*, in which the traditional assumption of independent relevance of documents with respect to a query does not hold. Here the *quality* of the retrieval system cannot be quantified as an aggregation of *quality* of each retrieved document, but as a quality of the whole set of retrieved documents.

TREC has also acknowledged the issue and, since 2009, has started a diversity task in its so-called Web tracks (Clarke, Craswell, & Soboroff, 2009). In the TREC 2009

Web Track, each of the 50 different topics was structured into a representative set of subtopics (sets of possible interpretations or aspects) related to different user needs. Topics were categorized as ambiguous or faceted, depending on whether its subtopics are interpretations or aspects of the query. For examples of both topic types, see Figure 5. Also, subtopics are classified as *navigational* or *informational*.

```
<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 5. Examples of ambiguous and faceted queries of TREC 2009 Web track topics.

It is important to stress that in most situtations the subtopics are not known by the TREC competition participants –or by systems being tested in research experiments using the datasets– when retrieving documents, and they are only used for evaluating the systems' output. This means that systems targetting diversity in their results must obtain the possible subtopics of a topic from their own sources or develop any other

way to promote diversity. For example, Agrawal, Gollapudi, Halverson, & Ieong (2009) use the Open Directory Project (ODP) categories, Santos, Macdonald, & Ounis (WWW, 2010) exploit query reformulations from commercial search engines (using their public APIs) to identify those subtopics and, in a very different approach, others such as Zhai, Cohen, & Lafferty (2003) use language models with KL-divergence or simple mixture models to calculate document similarity to increase the dissimilarity between documents of a result list, thus aiming to cover as many subtopics as possible.

2.2.1 Diversity and Novelty Metrics

2.2.1.1 Subtopic Retrieval Metrics

Zhai, Cohen, & Lafferty (2003) present an initial 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 the first *K* results, the retrieved proportion the n possible subtopics

$$S\text{-}recall@K = \frac{\left|\bigcup_{i=1}^{K} subtopics(d_i)\right|}{n}$$

where $subtopics(d_i)$ is the set of subtopics covered by document d_i. 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:

$$S\text{-}precision@r = \frac{\text{minRank}(L_{opt}, r)}{\text{minRank}(L, r)}$$

where L is the ranked list of retrieved documents, L_{opt} is an optimal system for minRank, and:

$$\text{minRank}(L, r) = \min\{K : S\text{-}recall@K \geq r\}$$

These two metrics do not take into account that redundancy (retrieving many documents covering the same subtopic) may not be desirable. For this purpose, a cost function that sums the cost b of presenting new documents and the cost a of presenting single subtopics is proposed:

$$cost(d_1, \ldots, d_K) = a\sum_{i=1}^{K} |subtopics(d_i)| + Kb$$

Then, analogously to S-precision, a *weighted subtopic precision* (WS-precision) is defined:

$$WS\text{-}precision@r = \frac{\text{minCost}(L_{opt}, r)}{\text{minCost}(L, r)}$$

This metric generalizes S-precision in that the latter is WS-precision with $b = 1$ and $a = 0$.

In (Chen & 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:

$$S\text{-}recall@K = \frac{\left|\bigcup_{i=1}^{K} subtopics(d_i)\right|}{n} = \frac{1}{n}\sum_{s} 1\text{-}call_s@K$$

2.2.1.2 Redundancy-Penalization Metrics

Clarke, et al. (2008) stress the fact that most IR 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 (incompleteness, 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 R_k of the k-th document for a user need u should be considered in the light of documents ranked above *k*:

$$P(R_k \mid u, d_0, \ldots, d_{k-1})$$

In Clarke's approach, the information need and the documents are modeled in a space of *information nuggets* $\mathcal{N} = \{n_1, \ldots, n_m\}$ so that user needs $u \subset \mathcal{N}$ and documents $d \subset \mathcal{N}$ are represented as "bags of nuggets". A *nugget* in this approach is an abstraction intended to stand for an indivisible unit of meaning, which can be materialized in different ways, always representing a binary property about documents. Assuming mutual independence of *nuggets* belonging to a document or a user need the probability of relevance can be estimated as:

$$P(R_k \mid u, d_0, \ldots., d_{k-1}) = 1 - \prod_{i=1}^{m}\left(1 - P(n_i \in u)\, P(n_i \in d_k) \prod_{j=0}^{k-1} P\left(n_i \notin d_j\right)\right)$$

On one hand, to estimate $P(n_i \in d)$, the authors assume that a human assessor (denoted by a binary function $J: \mathcal{D} \times \mathcal{N} \rightarrow \{0,1\}$) always can determine negative judgements ($J(d, n_i) = 0$) without error and positive ones ($J(d, n_i) = 1$) with some probability α. On the other hand, estimating $P(n_i \in u)$ would require knowledge of user probabilities, which is not always available. In such frequent cases, the authors propose the assumption of a fixed, independent probability $P(n_i \in u) = \gamma$. Based on all this, the formula can be further simplified to:

$$P(R_k \mid u, d_0, \ldots., d_{k-1}) = 1 - \prod_{i=1}^{m}\left(1 - \gamma\, \alpha\, J(d_k, n_i) \prod_{j=0}^{k-1}(1-\alpha)\right)$$
$$\approx \gamma\, \alpha \sum_{i=1}^{m} J(d, n_i)\, (1-\alpha)^{r_{i,k-1}}$$

where $r_{i,k-1} = \sum_{j=0}^{k-1} J\left(d_j, n_i\right)$. Since $\gamma\, \alpha$ is constant, it has no relative impact when comparing systems (as the metric is intended to), whereby the authors define a gain function for their nDCG-based metric as:

$$G(k) = \sum_{i=1}^{m} J(d, n_i)\, (1-\alpha)^{r_{i,k-1}}$$

The resulting metric is α-nDCG:

$$\alpha\text{-}nDCG = \frac{1}{\alpha\text{-}IDCG} \sum_{k=1}^{n} \frac{G(k)}{\log_2(k+2)}$$

where α-IDCG is the ideal or maximum value of α-DCG, which for practical purposes is usually calculated approximately with a greedy approach.

Clarke has continued studying possible unifications of α-nDCG and others. In (Clarke C. , Craswell, Soboroff, & Ashkan, 2011) a unified framework for novelty and diversity metrics is proposed. Diversity is accommodated through a linear combination of measures computed on individual subtopics (see the description of intent-aware metrics of the next subsection). Novelty is accommodated by penalizing redundancy. In fact, some of th already presented metrics can be explained under this framework. After conducting some experiments with the test collection of the TREC 2009 Web track, results indicate that these metrics work as intended.

2.2.1.3 Intent-Aware Metrics

Agrawal, Gollapudi, Halverson, & Ieong (2009) propose a generalization of standard IR metrics to acknowledge the possible intents of a query. Hence, given a metric M –such as nDCG, MRR, MAP, ERR–, its intent-aware version IA-M is defined as:

$$IA\text{-}M(q) = \sum_{c} p(c|q)\, M(q|c)$$

Here, $M(q|c)$ means a modification of M in which the documents that do not belong to category c are considered not relevant, and those that belong to the category will keep their relevance score. For example, the intent-aware version of ERR is:

$$IA\text{-}ERR(q) = \sum_{c} p(c|q) \sum_{k=1}^{N} \frac{1}{k} \prod_{i=1}^{k-1} (1 - r_i^c)\, r_k^c$$

2.2.2 Diversification Methods

In the last section, evaluation metrics have been introduced to measure the effectiveness of systems at the novelty and diversity task, but these measures have worst-case NP-hard computation time (Carterette, Information Retrieval, 2011). The primary consequence of this is that there is no ranking principle akin to the PRP for document relevance that provides uniform instruction on how to rank documents for novelty and diversity. Therefore alternative approaches must be applied to optimize system towards novelty and diversity. In particular, *greedy reranking techniques* have been widely used. Here we review some of these techniques and other solutions.

2.2.2.1 Maximal Marginal Relevance

One of the first references on diversity in IR appears in (Carbonell & Goldstein, 1998), where a method for combining query relevance and the so called *information novelty* for text retrieval is presented. This kind of method is appropriate for scenarios where there is a considerable big set of relevant documents, in which information redundancy is often observed.

Specifically, the Maximal Marginal Relevance (MMR) criterion establishes 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, a document so that is

$$\underset{d_i \in R \setminus S}{\operatorname{argmax}} \left[\lambda \operatorname{rel}(d_i, q) - (1 - \lambda) \max_{d_j \in S} \operatorname{sim}(d_i, d_j) \right]$$

where λ is a parameter taking values between 0 an 1, q is the query for which documents are retrieved, d_i and d_j are documents from the document collection R, S represents the set of higher-ranked, retrieved documents and rel and sim are functions of document-query relevance and document-document similarity, respectively.

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 MRR 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 IR and RS.

2.2.2.2 IA-Select

In (Agrawal, Gollapudi, Halverson, & Ieong, 2009), the authors assume that there is a taxonomy of information whose topical level models the user intents, 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, the authors develop an objective that tradeoffs 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 satisfying the user intent represented by the category the query belongs to, they pose the DIVERSIFY(K) objective:

$$\text{DIVERSIFY(K)} = \underset{S \subset R(q):|S|=k}{\operatorname{argmax}} P(S|q)$$

The probability $P(S|q)$ will represent the probability that the set of documents S satisfies the average user by averaging, for each category c representing a possible intent for the query q, the probability that some documents of S satisfy the category c:

$$P(S|q) = \sum_c p(c|q) \left(1 - \prod_{d \in S} \left(1 - V(d|q,c) \right) \right)$$

Since DIVERSIFY(K) is NP-hard and its solutions is not necessarily unique, a more efficient solution for determining a good solution for the problem is used here. Particularly, the so-called *IA-Select* (*intent-aware select)* algorithm is a greedy algorithm that selects from a set of retrieved documents the document that maximizes:

$$IA\text{-}Select(d|S,q) = \sum_c p(c|q)\, V(d|q,c) \prod_{d' \in S} \left(1 - V(d'|q,c) \right)$$

The set of documents S contains the documents selected by IA-Select in the previous steps. Since $P(S|q)$ is submodular, IA-Select has the nice property of finding a solution S' to that $P(S'|q) \geq \left(1 - {}^{1}/_{e}\right) P(S^*|q)$, where S^* is one of the optimal solutions.

2.2.2.3 Learning to Rank Approaches

Recently in the IR field there has been a significant growth of the application of Machine Learning approaches, which has received the name of *Learning to Rank*. The problem in diversity in IR has also been approached from this point of view. The first

reference found is (Radlinski, Kleinberg, & Joachims, 2008), where two different algorithms, *Ranked Explore and Commit* and *Ranked Bandits Algorithm*, use data of user clicks to produce diverse rankings. Yue & Joachims (2008) also present an *Learning to Rank* approach for learning diverse subsets using structural SVMs. More recently, Slivkins, Radlinski, & Gollapudi (2010) present a scalable approach that takes into account document similarity and context whith appropiate theoretical foundations.

2.2.2.4 Portfolio Theory

Wang & Zhu (2009) studied the problem of ranking under uncertainty using Modern Portfolio Theory. While the classic PRP approaches deal with maximizing the effectiveness in ranked lists, it does not consider the implicit risk (measured as the variance of the overall effectiveness) that a given ranking may have. If the releance r of each document is considered a random variable, the expected value and the risk (variance) of the overall relevance of a ranked list are given by:

$$E[R_n] = \sum_{i=1}^{n} w_i \, E[r_i]$$

$$Var(R_n) = \sum_{i=1}^{n} \sum_{j=1}^{n} w_i \, w_j \, \sigma_i \, \sigma_j \, \rho_{i,j}$$

where w_i is the weight associated at the i-th position in the ranking, $E[r_i]$ is its expected relevance, σ_i is the standard deviation of the relevance and $\rho_{i,j}$ is the correlation coefficient between the relevances of documents at positions i and j. The authors focus on maximizing the following objective function:

$$O_n = E[R_n] - b \, Var(R_n)$$

For this purpose, a greedy approach is proposed so that, for each step $k \in \{2, \dots, n\}$ the following quantity is maximized:

$$E[r_k] - b \, w_k \, \sigma_k^2 - 2b \sum_{i=1}^{k-1} w_i \, \sigma_i \, \sigma_k \, \rho_{i,k}$$

In the paper, the authors show that this approach can improve the results for subtopic text retrieval of PRP and MMR approaches in terms of S-recall and other diversity metrics.

2.2.2.5 xQuAD

In (Santos, Macdonald, & Ounis, WWW, 2010) a novel algorithm for diversification is presented. The *xQuAD* (*explicit query aspect diversification*) algorithm makes use of query reformulation provided by commercial web search engines to derive new sub-queries that will cover the possible aspects of the initial query. So, given an ambiguous query and a ranking of retrieved documents R, xQuAD will greedily select and inserting in a new ranking S the document d in $R \setminus S$ maximizing the following mixture probability:

$$(1 - \lambda) \, P(d|q) + \lambda \, P(d, \bar{S}|q)$$

where $P(d|q)$ is the probability of the document d being observed given the initial query q and $P(d, \bar{S}|q)$ the probability of observing the document but not the documents already in S. Using the subset $\{q_i\}$ of queries and some simplifying assumtions, one can expand $P(d, \bar{S}|q)$ so the objective formula for xQuAD becomes:

$$(1-\lambda)\,P(d|q)+\lambda\sum_{q_i}\left[P(q_i|q)\,P(d|q_i)\prod_{d_j\in S}\left(1-P(d_j|q_i)\right)\right]$$

More recently, the same authors (CIKM, 2010) proposed a way to determine a way of selecting λ optimally for each query, adapting the specific need for diversification.

2.2.2.6 Intent Hypothesis

Hu, Zhang, Chen, & Wang (2011) found that current click models based on the *examination hypothesis* cannot fully explain user clicks by relevance and position bias. This examination hypothesis states that an document in a result list has been clicked if and only if it was examined and was relevant. They deduce that there is a intent bias derived from the relation between the user need (intent) and the submitted query in every search session, so the examination hypothesis needs to be redesigned to incorporate this intent bias. In particular, the *intent hypothesis* is based on three premises:

- The user clicks a document if and only if it is examined and needed by the user.
- If a document is irrelevant, the user will not need it.
- If a document is relevant, whether it is needed is only influenced by the gap between the user's intent and the query (intent bias).

Previous state of the art click models can be easily modified to adjust to the new itent hypothesis.

2.2.2.7 DivRank

A diversity-aware alternative for *PageRank* called *DivRank* is presented in (Mei & Guo, 2010). As PageRank, DivRank is based on a random walk over a network of linked documents with a teleportation component and supposes that connected documents tend to be more similar than others whose linkage is weaker. The particularity of DivRank is that the transition probabilities from document d_i to d_j are adjusted at each step of the random walk to be proportional to the number of times the document d_j 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.

2.3 Novelty and Diversity in Recommender Systems

2.3.1 Overview

Similarly to IR, research in RS has been strongly focused on achieving accuracy in matching user interests, either in terms of rating value prediction error (as measured by MAE or RMSE, see Adomavicius & Tuzhilin, 2005) or in terms of ranking quality (as measured by IR metrics). Nevertheless, evaluating recommender systems only this way may present serveral considerable limitations. In (Herlocker, Konstan, Terveen, & Riedl, 2004) it is stated that "*there are properties different from accuracy that have a larger effect on user satisfaction and performance*" such as coverage or non-obviousness. In (McNee, Riedl, & Konstan, 2006) the authors propose a new-user centric direction for evaluating recommender systems based on three aspects: diversity, novelty and the user needs and expectations in a recommender. We analyze here the two first aspects.

Novelty and diversity are highly desirable features for automatic recommendation. In most scenarios, the purpose of recommendation is inherently linked to a notion of discovery, as recommendation makes most sense when it exposes the user to a relevant experience that she would not have found by herself –obvious, however accurate recommendations are generally of little use. Besides, user interest prediction involves inherent uncertainty, since it is based on implicit, incomplete evidence of interests, where the latter are moreover subject to change. Therefore, avoiding a too narrow array of choice is generally a good approach to enhance the chances that the user is pleased by at least some recommended item. Sales diversity may enhance businesses as well, leveraging revenues from market niches (Fleder & Hosanagar, 2009).

Reported contributions in this area involve the definition of algorithms and strategies to enhance novelty and diversity, as well as methodologies and metrics to assess how well this is achieved. From the common understanding that novelty and diversity play a fundamental part as dimensions of recommendation utility, most authors have dealt with these properties as opposing goals to accuracy, stating the problem as a multi-objective optimization issue, where an optimal trade-off between accuracy and diversity is sought.

Novelty and diversity are different though related notions. The novelty of a piece of information generally refers to how different it is with respect to "what has been previously seen or known", by a specific user, or by a community as a whole. Some authors even make a distinction between novelty and *serendipity* (McNee, Riedl, & Konstan, 2006 and Herlocker, Konstan, Terveen, & Riedl, 2004). Serendipity is defined as the quality of novel items that would not have been able to be discovered without the help of the recommendation system.

Novelty in recommendation is specially relevant to exploit the *Long Tail effect*, i.e., the situation where a few items are extremely popular and there is the rest of them are much less known (Figure 6). As estated by Anderson (2006), recommender systems may benefit from *selling less of more*, that is, recommending less wide-known items to more users instead of focusing on hihgly-popular items.

Diversity generally applies to a set of items, and is related to how different the items are with respect to each other. This is related to novelty in that when a set is diverse, each item is "novel" with respect to the rest of the set. Moreover, a system that promotes novel results tends to generate diverse results for each user over time and also enhances the global "diversity of sales" from the system perspective. It is worth to make a distinction between *individual diversity* and *aggregate diversity*. The first case accounts for how different are items in a recommendation list for a only user, which is normally the notion of diversity employed in most works. Nevertheless, aggregate diversity –understood as the total amount of different items a recommendation algorithm can provide to the community of users– (Adomavicius & Kwon, to appear) is also a very interesting quality of a RS as a whole.

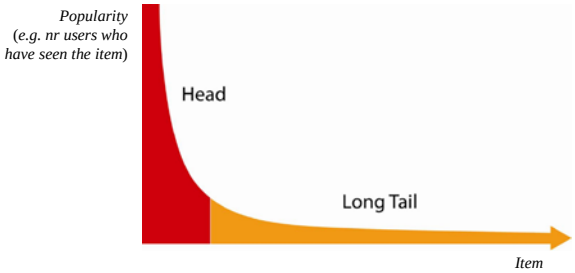

Figure 6. The long tail effect

2.3.2 Topic Diversification and Intra-list Similarity

A common specific definition of diversity in the literature is the average pairwise dissimilarity between recommended items. Using this notion, Ziegler, McNee, Konstan, & Lausen (2005) introduce the *topic diversification* method to balance and diversify personalized recommendations list in order to reflect the user's complete spectrum of interests. This article regards the recommendation lists as entities on their own, rather than pure aggregations of single items.

For evaluation purposes, the authors measure the diversity of a recommendation list by a metric called *intra-list similarity* (ILS). Given a similarity function $\text{sim}: \mathcal{I} \times \mathcal{I} \rightarrow [-1, +1]$ the ILS for a recommendation list R is defined as:

$$ILS(R) = \frac{1}{2} \sum_{i \in R} \sum_{j \in R} \text{sim}(i, j)$$

Note that ILS is permutation-invariant for the elements of R, that is, it is insensitive to the order of recommended items. This can be a considerable limitation as far as users do not necessarily browse down to the end of the list, whereby the order in which items are presented may heavily influence the pracitcal utility of the recommendation. We address this limitation as part of our research, as we report in chapter

The *topic diversification* algorithm reranks a recommendation list R. It needs a set similarity metric $\text{sim}^*: \mathcal{P}(\mathcal{I}) \times \mathcal{P}(\mathcal{I}) \rightarrow [-1, +1]$ that can be extracted using, e.g., a classification taxonomy of features of the items. Basically, the algorithm consists in choosing greedily the item i that minimizes

$$\lambda \, rank_R(i) + (1 - \lambda) \, rank_{\text{sim}^*(S,\{\cdot\})}(i)$$

where $rank_R$ returns the position in the original list of the item i and $rank_{\text{sim}^*(S,\{\cdot\})}$ returns the position in the rank created by sorting the elements of $R \setminus S$ by their similarity to the items in S in decreasing order. The authors suggest that their algorithm ressembles the membrane's selective permeability of molecular biology. This approach is also very similar to the Maximal Marginal Relevance scheme proposed in Information Retrieval (IR) for search diversification and automatic summarization (Carbonell & Goldstein, 1998).

The authors conducted both offline and online experiments on a book exchanging community. The results in the offline experiment suggest that topic diversification may reduce the accuracy of recommendations in terms of precision an recall, but would reduce significantly the average ILS, thus incrementing the diversity of results.

2.3.3 Diversity as a Quadratic Optimization Problem

In (Zhang & Hurley, 2008) the problem of diversity is posed as a joint maximization problem of two objective functions reflecting preference similarity and item diversity with constraints. They bring intra-list diversity to a more formal formulation and problem statement, as follows. Let $\mathbf{y} \in \{0,1\}^{|\mathcal{I}|}$ be the vector indicating a top-N recommendation (where where y_k is 1 in case document d_k is selected and 0 otherwise, so $\mathbf{1}^{\mathrm{t}} \, \mathbf{y} = N$), $D \in \mathbb{R}^{|\mathcal{I}|,|\mathcal{I}|}$ the matrix containing the distances between items and $\mathbf{m}_u$ a vector indicating the relevance of each item for user u, then the diversity of a recommendation and its cumulative gain can be computed in terms of $\mathbf{y}$:

$$ILD = \mathbf{y}^t D \mathbf{y} \text{ and } CG = m_u^t \mathbf{y}$$

Consequently, the problem of diversification can be formally defined as finding $\mathbf{y}^*$ such that

$$\mathbf{y}^* = \underset{\substack{\mathbf{y}\in\{0,1\}^{|\mathcal{I}|} \\ \mathbf{1}^t\,\mathbf{y}=N}}{\arg\max}\,(1-\theta)\;\alpha\;\mathbf{y}^t D\mathbf{y} + \theta\;\beta\;m_u^t\mathbf{y}$$

where θ is a trade-off parameter between diversity and accuracy, and α and β are normalization parameters so both components lie in the interval [0,1]. To solve this complex problem, the authors propose to relax it into a real-valued problem. Then, the real-valued problem is solved by linear and quadratic programming algorithms which are much easier to solve that the discrete-valued case. Finally, they quantize the values of the real-valued solution to a candidate binary solution $\mathbf{y}^*$.

The authors introduce an interesting evaluation approach consisting of the biased selection of novel test items, whereby evaluating for novelty is achieved by studying the accuracy on such difficult items.

2.3.4 Popularity, Long-tail Items and Recommendation Algorithms

Zhou, Kuscsik, Liu, Medo, Wakeling, & Zhang (2010) propose some other ways to assess diversity and novelty. For assessing system-wide diversity, the personalization of a recommender system is the average over all pairs u, v of users of the distance between their top-N recommendation lists R_u and R_v is defined as

$$h_{u,v} = 1 - \frac{|R_u \cap R_v|}{N}$$

Averaged over all pairs of users, this metric h should evaluate the capacity of the system to provide user-specific recommendations.

As a measure of surprisal or novelty, they propose *mean self-information* (MSI), which computes the mean of the unexpectedness of each item $i \in R$ relative to its global popularity:

$$MSI(R) = \frac{1}{|R|}\sum_{i\in R}\log\frac{|\mathcal{U}|}{|\{u \in U | i \in \mathbf{u}\}|}$$

The authors propose algorithms which target both metrics, by means of hybrid strategies combining collaborative filtering with graph spreading techniques.

Celma & Herrera (2008) take 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 analyze the paths leading from recommendations to the long tail through similarity links. Specifically the analize collaborative filtering (CF) and content-based (CB) recommendations. For the case of CF recommendations, the topology of the item similarity network leads to poor discovery ratio. On the other hand, CB recommendations can provide more novel recommendations with lower perceived quality. Solutions suggested include promoting unknown artist of the long tail of the popularity distribution or selecting CF or CB depending on the users's needs.

2.3.5 Temporal Diversity

Lathia, Hailes, Capra, & Amatriain (2010) deal with temporal diversity in CF recommender systems. In a realistic scenario, users interact with recommender systems iteratively over time, so new models must be trained regularly to adapt to new users, new items or updated user profiles. They carried out two experiments. An online

experiment showed that user's perception of the recommendations lists degrades if the do not show diversity with respect to paste recommendations to the same user. The offline experiment compared the temporal diversity of some CF recommenders among time, reaching interesting conclusions:

1) Item-based recommenders have on average more temporal diversity than matrix-factorization ones.
2) As users' profiles increase, temporal diversity decreases.
3) The more a user interacts (rates) with the system in a session, the more diverse the next recommendations will be.
4) 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 diversity.

One may draw from these observations the conclusion that improving temporal diversity is an important task. In the paper two methods are proposed:

1) Switching between recommenders, so it is easier that the recommended items are not be the same for each list over time while mantaining some accuracy. The swiching period can be fixed or user-specific.
2) Randomly reranking recommendation lists so a specific amount of top-N recommendations are replaced with others of lower predicted preference but more diverse with respect to previous recommendations.

2.3.6 Aggregate Diversity

Adomavicius & Kwon (to appear) address diversity as the ability of a system to recommend as many different items as possible over the whole population. This form of *aggregate diversity* is measured as the size of the set of all items a recommender system is able to recommend to its users as a whole:

$$aggr\text{-}div = \left| \bigcup_{u \in \mathcal{U}} R_u \right|$$

As a diversity-enhancing approach, They propose a parametric reranking method combining standard CF recommenders with other ranking criteria that promote aggregate diversity but have poor accuracy, so they compensate. If $rank_{CF}(i) = 1/\hat{r}(u,i)$ denotes the function that defines the ranking R in ascending order, and $rank_X(i)$ is one of the alternative aforementioned ranking criteria, then the proposed reranking is defined through a ranking threshold τ in the following manner:

$$rank_X(i;\tau) = \begin{cases} rank_X(i) & \text{if } \hat{r}(u,i) \geq \tau \\ \alpha_u + rank_{CF}(u,i) & \text{if } \hat{r}(u,i) < \tau \end{cases}$$

$$\text{where } \alpha_u = \max_{i:\hat{r}(u,i)\geq\tau} rank_X(i)$$

With this reranking, all items with predicted rating above τ will be first included in the ranking sorted by the $rank_X$ criterium, and those below the threshold will maintain their relative sorting, but will appear in lower positions than those of the first group. This way, by selecting τ appropiately it is possible to maintain part of the original accuracy while maximizing the aggregate diversity. As possible implementations of $rank_X$ the authors propose the following:

- Inverse popularity: $rank_{InvItemPop}(i) = |\{u \in \mathcal{U} \mid i \in \mathbf{u}\}|$
- Reverse rating: $rank_{RevRating}(i) = \hat{r}(u,i)$

- Average rating: $rank_{AvgRating}(i) = \text{avg}_{u':i\in\mathbf{u}'}\,\hat{r}(u',i)$
- Item rating variance: $rank_{ItemVar}(i) = Var_{u:i\in\mathbf{u}}\big(\hat{r}(u,i)\big)$

2.3.7 User Profile Partitioning

Zhang & Hurley (2009) apply clustering techniques for recommending novel items. In standard situations, the recommendation is based on aggregate similarity metrics of items to the user profile, so the influence of novel items is lost in the aggregation. The authors propose to partition the user profile into clusters and compose a recommendation list of items that matches well with each cluster (Figure 7).

The steps to produce a recommendation for a user u are:

1) Partition the items in the user profile $\mathbf{u}$ into M clusters $\{C_k^u\}_{k=1}^M$ so the intra-cluster distance is minimized. The clustering or partitioning strategies suggested are graph partitioning, *k-means* and modularity maximization.
2) For each cluster, generate a recommendation taking it as a whole user profile. The proposed approach here was taking the $h = \max(N, M)$ clusters with higher aggregate novelty and generate for those selected a top-N_k recommendation where $N_k = \lfloor N/h \rfloor$.
3) Aggregate the recommendations for each cluster together to form a unique recommendation.

Additionally, the authors carried out a dimension reduction strategy based on matrix factorization to help uncover similarities that are otherwise difficult to recognize in a higher-dimensional space.

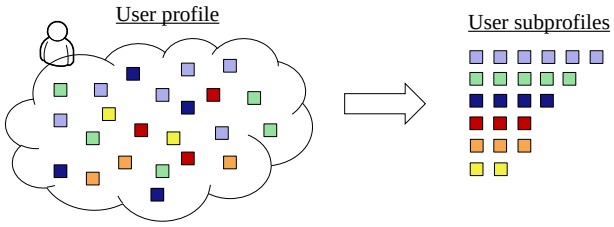

Figure 7. User profile partitioning

2.3.8 Information Theoretical Metrics for Diversity and Novelty

In (Bellogín, Cantador, & Castells, 2010) a study of heterogeneity in music recommendations is conducted. They propose their own metrics for assessing diversity, relative diversity and novelty based in Information Theory concepts such as entropy or mutual information. Given a user u, the diversity of a recommendation list R_u with respect to a set $\mathcal{R}_u$ of recommendations defined as:

$$div(R_u) = -\sum_{i\in R_u} p_{u,i}\log p_{u,i} \quad \text{where} \quad p_{u,i} = \frac{|\{R'_u \in \mathcal{R}_u | i \in R'_u\}|}{|\mathcal{R}_u|}$$

Relative diversity deals with differences between two recommendations and is defined as

$$div(R_u, S_u) = \sum_{i\in R_u\cap S_u} p_{R_u,i}\log\frac{p_{R_u,i}}{p_{S_u,i}} \quad \text{where} \quad p_{R_u,i} = \frac{[i \in R_u]}{|R_u|}$$

Finally, they define novelty as

$$nov(R_u) = -\sum_{i \in R_u} p_i \log p_i \quad \text{where} \quad p_i = \frac{|\{u' \in \mathcal{U} | i \in R_{u'}\}|}{|\mathcal{U}|}$$

3. Information Retrieval Diversity for Recommender Systems

In this chapter we explore the adaptation of diversity metrics, techniques, and principles from ad-hoc IR to the recommendation task. Particularly, we introduce the concept of *aspect space* as a mean to translate two key notions of IR diversity, document similarity and query intents, to their correspondences to the RS field: *item similarity* and *user profile aspects*, respectivelty. We propose ways of modeling aspect spaces using explicit and implicit information available from collected data, being the implicit case especially interesting in cases where the available information is limited to collaborative filtering data. Empirical results support the proposed approaches and provide further insights.

The contents of this chapter have been published in (Vargas, Castells, & Vallet, SIGIR, 2011).

3.1 Introduction

In general terms, and most particularly in common practical scenarios, recommendation can be seen as an IR task. Interestingly, the diversity issue has been stated and addressed quite differently in the research on the topic so far in RS and ad-hoc IR respectively. In particular, diversity has been studied under a quite specific motivation and precise problem definition(s) in the IR community –building around the problem of uncertainty in user queries– along with formally grounded and well understood diversity metrics, with a theoretical depth and a drive towards standardization (backed by a specific TREC diversity task) which are not presently found or equally emphasized in the RS literature on the topic. 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 RS diversity might benefit from the insights and ongoing progress in search diversity –and vice-versa.

In this chapter we explore the adaptation of diversity models, metrics, and methods from ad-hoc IR into a RS setting. Specifically, we propose the notions of *item similarity* and *user profile aspects* as analogues of document similarity and query intents, respectively, upon which we adapt IR diversity techniques and methodologies to a recommendation task. We consider two scenarios that differ in the available information for the construction of a item similarity function and a user aspect space used by our RS-adapted diversification methods and diversity metrics. In one scenario, we propose an approach for the extraction of item features and user aspects based on latent factors when the only available information relates to the interaction between users and items.

3.2 Recommendation Diversity vs. Search Diversity

Diversity in RS is generally motivated as a means to reduce redundancy under the assumption that recommending too similar items is less profitable for the user –and the vendor– than offering a more varied experience. The rationale for diversity is often

stated in association with the notion of novelty and surprisal, upon the understanding that recommendation value is to a significant extent related to discovery in the user experience. Looking back for a connection to diversity in ad-hoc IR, one finds that the issues of ambiguity and underspecification are generally absent from the problem statement in the RS 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 RS diversity. So does the principle of diversification as a means to minimize the risk of underperformance extremes, which is also common in the IR literature (Agrawal, Gollapudi, Halverson, & Ieong, 2009).

Query ambiguity and underspecification are modeled in terms of query interpretations, categories, aspects, nuggets, subtopics, and similar elements in ad-hoc IR. An analogy can be drawn in the RS setting by considering an equivalent notion of *user profile 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 preference 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.

If one is able to give a consistent approximation to item similarity and user profile aspects in the context of a RS, the theories and metrics in search diversity could be adapted to the recommendation task. This would bring benefits such as

a) a new perspective and rationale for diversity in RS in terms of theory and models, and
b) new diversity metrics for RS, such as the intent-aware metrics (Agrawal, Gollapudi, Halverson, & Ieong, 2009) or α-nDCG (Clarke, et al., 2008).

Additionally, such metrics would bring in several important properties currently lacking in RS diversity studies:

a) the introduction of metrics that take into account the order of items when measuring the overall recommendation diversity (i.e. top positions are more important);
b) the consideration of diversity only in the presence of relevance;
c) related to this, the assessment of accuracy and diversity altogether by a single metric; and
d) a step towards a shared consensus on common metrics and methodologies.

3.3 The Concept of Aspect Space

We consider a set of aspects $a \in \mathcal{A}$ which model, for all users, their different and disjoint interests derived from their profiles. The idea behind these aspects could be expressed as *the more aspects of the user profile are covered, the more diverse the recommendation will be perceived by the user*. As these interests are not always equally representative of the user profile, we find it convenient to represent the user profile aspect space as a probability distribution of the aspects for the user profile, that will be denoted as $p(a|u)$. Further, we will suppose that we have full information about each aspect space (meaning $\sum_a p(a|u) = 1$).

Following the analogy of query intents, it is necessary a way to determine the extent the user aspects are covered by the items of the collection. Same as before, we will consider for each item an aspect space defined by a probability distribution for the aspects denoted as $p(a|i)$ with full information ($\sum_a p(a|i) = 1$). Given these item aspect spaces, it is also possible to define a similarity metric between items:

$$\text{sim}: \mathcal{I} \times \mathcal{I} \rightarrow [0,1]$$

Finally, we will say that a certain aspect belongs to the user profile aspect of u or the item aspect space of i if $p(a|u) \neq 0$ or $p(a|i) \neq 0$, respectively.

Before showing how we create an aspect space, let us see first their application for adapting diversification algorithms and diversity metrics.

3.4 Adapted Aspect-Based Diversification Algorithms

Given that the diversity problem is generally stated as a NP-hard problem (Carterette, Information Retrieval, 2011), it is common to solve it applying a greedy algorithm where a baseline ranking R is diversified into a re-ranked list S by iteratively picking the item $i \in R \setminus S$ which maximizes an objective function. Specifically, we adapt here two well-known algorithms from search diversity: IA-Select (Agrawal, Gollapudi, Halverson, & Ieong, 2009) and MMR (Carbonell & Goldstein, 1998).

In the IA-Select scheme, the objective function is defined as:

$$\sum_c p(c|q)\, V(d|q,c) \prod_{d' \in S} \big(1 - V(d'|q,c)\big)$$

In our RS context, we translate the taxonomy of categories c to a set of aspects $a \in \mathcal{A}$ modeling the user profile aspects and items aspects. In our adaptation of the objective function, we consider that $V(i|u,a) = \hat{r}(u,i)p(a|i)$. Finally, the objetive function for IA-Select in RS is defined as:

$$\sum_{a \in \mathcal{A}} p(a|u)\hat{r}_{\text{norm}}(u,i)p(a|i) \prod_{j \in S} \big(1 - p(a|j)\hat{r}_{\text{norm}}(u,j)\big)$$

In MMR for search diversity the objective function to maximize has the following formulation:

$$\lambda\, \text{rel}(d_i, q) - (1-\lambda) \max_{d_j \in S} \text{sim}\big(d_i, d_j\big)$$

in our RS context we adapt it in a straightforward way by translating the concept of document simmilarity to item similarity:

$$\lambda\, \hat{r}_{norm}(i,u) - (1-\lambda) \max_{j \in S} \text{sim}(i,j)$$

Note that this approach is very similar to that of Ziegler, McNee, Konstan, & Lausen (2005), but in this case we do not proceed with a rank normalization of components.

3.5 Adapted Diversity Metrics

To evaluate the quality of diversified recommendations we adapt measures such as the intent-aware metrics (Agrawal, Gollapudi, Halverson, & Ieong, 2009) and α-nDCG (Clarke, et al., 2008), where aspects play the role of categories (or subtopics), and user

profiles play the part of queries. That is, for instance, given a user u, the intent-aware nDCG of the recommendation to u is defined as

$$IA\text{-}nDCG = \sum_{a \in \mathcal{A}} p(a|u)\, nDCG(u|a)$$

where, as defined in (Agrawal, Gollapudi, Halverson, & Ieong, 2009), $nDCG(u|a)$ counts as relevant items only the ones that are relevant for u and have the aspect a. For the case of α-nDCG the adaptation results in the following modification of the gain function:

$$G(k) = \sum_{a} r(i_k, u)(1-\alpha)^{c_{a,k-1}} \text{ where } c_{a,k-1} = \sum_{l=0}^{k-1} \hat{r}(i_l, u; a)$$

where $r(i, u; a)$ is the user u preference for i in case the item has the aspect a, and 0 otherwise.

3.6 Aspect Space Extraction

In an ideal situation, we would have both user profile and item-specific information for modeling directly their associated aspect spaces. However, in many recommendation scenarios the information associated with the user profile is limited to a rating or access information of items in the collection, and item feature information may not be available, or may be incomplete.. Therefore it is necessary to create an aspect space from the available information, and we consider two different scenarios:

a) There is available information about item features (or attributes), such as genre, author, language, etc. This means that user profile aspects can be derived from the item features and the specific relation between the user and the items. We define this scenario as an *explicit aspect space.*
b) There is no information at all about the characteristics of the items. Therefore the user profile aspects can be solely modeled upon the relation between the users and documents. We define this scenario as an *implicit aspect space.*

3.6.1 Explicit Aspect Space Extraction

Let us consider the first scenario, in which there is an item feature space $\mathcal{F}$ so each item i has a set $\mathbf{i} \in \mathcal{F}$ of features. We shall assume that we have no a-priori information about how much each feature is representative of the item. In that case, we take the set of features $\mathcal{F}$ as our aspect set $\mathcal{A}$ and, for each feature f, the distribution probabilities for each item space will be defined as

$$p(f|i) = \frac{[f \in \mathbf{i}]}{|\mathbf{i}|}$$

For the distribution of user profile aspect spaces, we use the intuition that, the more items in the user profile contain a given feature, the more characteristic the feature is of the user interests:

$$p(f|u) = \frac{|\{i \in \mathbf{u} | f \in \mathbf{i}\}|}{\sum_{f' \in \mathcal{F}} |\{i \in \mathbf{u} | f' \in \mathbf{i}\}|}$$

Figure 8 illustrates the approach.

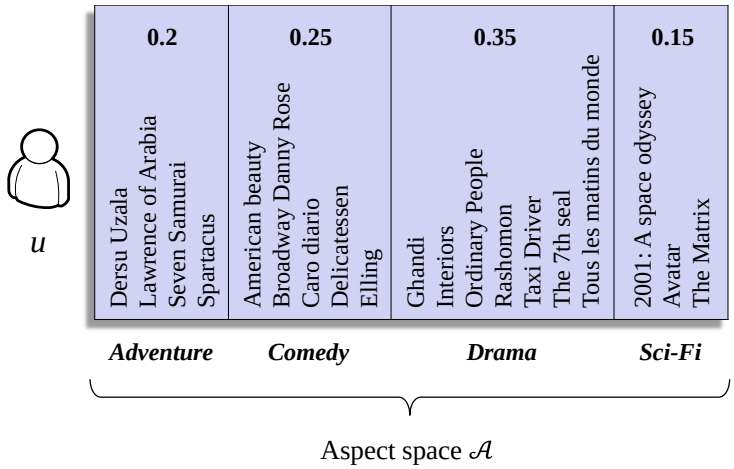

Figure 8. Explicit user profile aspects

Given this discrete aspect space, a similarity function between items can be defined using the cosine similarity for sets:

$$\text{sim}(i,j) = \frac{\sum_f [f \in \mathbf{i}][f \in \mathbf{j}]}{\sqrt{|\mathbf{i}||\mathbf{j}|}}$$

3.6.2 Implicit Aspect Space Extraction

This is a more challenging scenario, which is especially interesting for situations of pure collaborative filtering data, or situations where the content information about the items in the collection is missing or incomplete. Here, we need a way to extract implicit information about user interests derived from the interactions of users and items of our system. Drawing from principles of matrix factorization (Koren, Bell, & Volinksy, 2009), we use latent factor models as a way to create an aspect space for user profiles and items. In its most basic configuration, a matrix factorization technique considers a $\mathbb{R}^k$ factor space in which user profiles and items are represented by vectors p_u and q_i, respectively, so that the predicted preference values are determined by the inner product of the vectors:

$$\hat{r}(u,i) = q_i^t \, p_u$$

As suggested by the mentioned authors, the dimensions of the factor space can be interpreted as measures of the interests of the user. In our method, we identify these dimensions as the aspects of our probability space. Following the idea of the inner product for estimating preferences, it is natural to derive a similarity metric for items using the cosine between item factor vectors:

$$\text{sim}(i,j) = \frac{q_i^t \, q_j}{\sqrt{\|q_i\|\|q_j\|}}$$

For the estimation of the probability distribution for user profile and item aspect spaces, we carry out a binarization of the item feature vectors (given a $q_i \in \mathbb{R}^k$ vector and a constant $l < k$, we take the l dominant factors as the dimensions of the aspect space) to convert them into pseudo-explicit aspects. With this binarization process, the probability distributions can be used in the same way as for the explicit case. An illustration of this approach is shown is Figure 9.

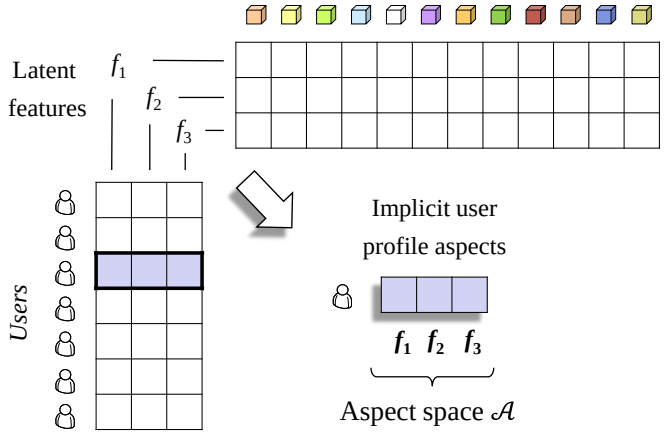

Figure 9. Implicit user profile aspects

3.7 Experiments

We have tested the behavior of the proposed approach on the MovieLens 100K dataset. We take as baselines two state of the art collaborative filtering algorithms: a common user-based nearest-neighbor (kNN) recommender, and a matrix factorization (MF) based algorithm (Koren, Bell, & Volinksy, 2009). We take the 80% training, 20% test data splits provided by the MovieLens distribution, with 5-fold cross-validation. For relevant judgments, we take as relevant (for each user) the items with a rating higher than 3 in the test set.

The adapted diversifiers (MMR and IA-Select) are used to re-rank the top 500 items returned by the baseline recommenders for each user. As aspect spaces, we test the two scenarios mentioned in section 3.6, one in which the diversifier uses the known item genre data, and one in which it extracts latent factors as the aspect space in the diversification algorithm, using rating information only.

Table 1 shows the performance of the different configurations using our adapted IR metrics, plus intra-list diversity (ILD) –based on the complement of item similarity in our aspect space on genres–, a common metric used in RS diversity (Ziegler, McNee, Konstan, & Lausen, 2005). Note that, for evaluation purposes, we have not considered the implicit aspect space since it is not as objective as the explicit aspects. It can be seen that the proposed diversification methods work properly, consistently improving the non-diversified baselines (bottom row). The IA-Select approach performs overall significantly better than the MMR scheme on the three IR metrics. We believe this is because it builds upon a common formalization of diversity as do the metrics (after Agrawal, Gollapudi, Halverson, & Ieong, 2009). Somewhat surprisingly, diversification with latent features performs better than with explicit ones for IA-Select on kNN, and MMR on both. We attribute this to the fact that latent features provide a more dense representation of items, and also more significant in terms of explaining the differences in interests between users, and the similarity between items. On the ILD metric, MMR and IA-Select perform similarly, and explicit features work clearly better than latent. This is probably because ILD ignores relevance –with respect to which IA-Select and latent features seem to do better.

		α-nDCG@50		ERR-IA@50		nDCG-IA@50		ILD@50	
		kNN	MF	kNN	MF	kNN	MF	kNN	MF
IA-Select	E	0.1589	**0.1838**	0.0409	0.0516	0.0604	**0.0755**	**0.8659**	0.8734
	L	**0.1596**	0.1597	**0.0465**	0.0458	**0.0618**	0.0637	0.7951	0.7817
MMR	E	0.1334	0.1652	*0.0367*	*0.0431*	0.0461	*0.0555*	0.8601	**0.8761**
	L	0.1320	0.1742	0.0373	**0.0528**	0.0492	0.0705	0.7906	0.7740
Baseline RS		0.1213	0.1451	0.0352	0.0425	0.0440	0.0561	0.7787	0.7655

Table 1. Four diversity metrics (α = 0.5 in α-nDCG) on different diversification approaches: MMR (with λ=0.5) and IA-Select, combined with explicit (E) and latent (L) features, on two baseline RS, based on kNN and MF respectively. The best value of each column is in bold. All differences to baseline are statistically significant ($p < 0.005$, Wilcoxon), except values in italics.

We have carried out additional experiments with further configurations, using different baseline recommender systems, and different metric cutoffs, the results from which also confirm our findings. The results were similarly positive with movie director as the explicit feature space. We plan to further explore the relation between the feature space and the effectiveness of diversification, under the intuition that the effectiveness should benefit from a higher dependency between features and user interests.

4. A Unified Metric Framework for Recommendation Novelty and Diversity Evaluation

Different evaluation metrics for novelty and diversity in Recommender Systems have been reported in the literature but the precise relation, distinction or equivalence between them has not been explicitly studied. Furthermore, the metrics reported so far miss important properties such as taking into consideration the ranking of recommended items, or whether items are relevant or not, when assessing the novelty and diversity of recommendations.

We present a formal framework for the definition of novelty and diversity metrics that unifies and generalizes several state of the art metrics. We identify three essential ground concepts at the roots of novelty and diversity: choice, discovery and relevance, upon which the framework is built. 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 of the browsing model, different metrics and variants unfold. We report experimental observations which validate and illustrate the properties of the proposed metrics.

The contents of this chapter have been published in (Vargas & Castells, 2011).

4.1 Introduction

While most research in the Recommender Systems has focused on accuracy in matching user interests, there is increasing consensus in the community that accuracy alone is not enough to assess the practical effectiveness and added-value of recommendations (Herlocker, Konstan, Terveen, & Riedl, 2004 and McNee, Riedl, & Konstan, 2006). In particular, novelty and diversity are being identified as key dimensions of recommendation utility in real scenarios, and a fundamental research direction to keep making progress in the field. Businesses are accounting for these aspects when engineering recommendation functionalities, and researchers have started to seek principled foundations for incorporating novelty and diversity in the recommendation models, algorithms, theories, and evaluation methodologies (Celma & Herrera, 2008, Fleder & Hosanagar, 2009, Zhang & Hurley, Recsys, 2008 and Ziegler, McNee, Konstan, & Lausen, 2005).

In this context, we identify the consolidation of a set of sound, well understood evaluation methodologies and metrics as a key issue to foster progress in this direction. Despite the raise of interest and work on the topic in recent years, we find that a clear common methodological and conceptual ground is still to be laid. Different evaluation metrics have been proposed in the literature but the relation, distinction or equivalence between them has not been explicitly studied. Furthermore, the metrics reported so far miss important properties such as taking into consideration the ranking of recommended items, or whether items are relevant or not, when assessing the novelty and diversity of recommendations. There is also variety in the principles and perspectives on which

different studies build, which would deserve analysis in order to better understand the potential connections and essential distinctions between them, fostering consensus and methodological convergence.

Our research aims to contribute to the identification of some of these connections and provide a formal ground for the unification of different ways to measure 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: choice, discovery and relevance, 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. We provide model estimation approaches on available observations of the interaction between users and items, thus providing for the practical computation of the metrics upon both explicit and implicit data. We report experimental observations validating and illustrating the properties of the proposed metrics.

4.2 Proposed Framework

The proposed metric framework is founded on three fundamental relations between users and items:

- *Discovery*: an item is seen by (or is familiar to) a user. We consider this fact independently from the degree of enjoyment / dislike, or whether the user consumed the item or not.
- *Choice*: an item is used, picked, selected, consumed, bought, etc., by a user.
- *Relevance*: an item is liked, useful, enjoyed, etc., by a user.

We model these three relations as binary random variables over the set of users and the set of items: $seen, choose, rel : \mathcal{U} \times \mathcal{I} \to \{0,1\}$. These three variables are naturally related: a chosen item must obviously be seen, and relevant items are more likely to be chosen than irrelevant ones. As a simplification, we assume relevant items are always chosen if they are seen (as illustrated in Figure 10), irrelevant items are never chosen, and items are discovered independently from their relevance. In terms of probability distribution, all these assumptions can be expressed as:

$$p(choose) \sim p(seen)p(rel) \tag{4.1}$$

where $choose$ is a shorthand for $choose = 1$, and same for the other two variables. Discovery, choice and relevance play different roles in our framework. Discovery is used as the basis to define item novelty models. Choice is used to build models of user browsing behavior over recommended lists of items. 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.

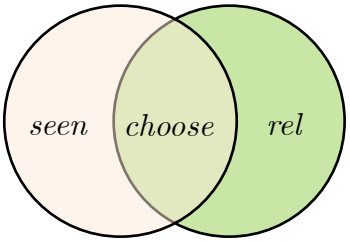

Figure 10. Discovery, choice and relevance models.

The starting point of the proposed framework is a general scheme where a recommendation metric is defined as the expected novelty of the recommended items the user will choose. Given a ranked list R of items recommended to a user u, this can be expressed as:

$$m(R|\theta) = C \sum_{i \in R} p(choose|i,u,R)\, nov(i|\theta) \qquad (4.2)$$

where C is a normalizing constant, and θ stands for a generic contextual variable which will allow for the consideration of different perspectives in the definition of novelty and diversity, as we will describe in the sections that follow. The metrics are thus determined by two main components: $p(choose|i,u,R)$, reflecting a browsing model grounded on item choice, as we shall see; and $nov(i|\theta)$, an item novelty model. In this scheme, the novelty or diversity of a recommendation is thus measured as the aggregate novelty of its constituent items. But the novelty of each item is considered *only* inasmuch as the user will actually want to use this item –as represented by $p(choose|i,u,R)$, denoting the probability that the target user u actually decides to use item i, when delivered within a recommendation R. This component provides a handle to make the metric sensitive to item relevance, and position in the ranking.

There are different ways in which the recommendation browsing model and item novelty can be developed. We describe them in detail in the next sections. For the time being, we intentionally denote the metric in formula 4.2 by a generic m, as it may reflect recommendation novelty or diversity depending on how the item novelty model, the browsing model, and θ are instantiated.

4.3 Item Novelty Models

Item novelty is the core element in the definition of recommendation novelty and diversity in our framework. Item novelty can be understood and defined in different ways, depending on which the resulting metrics differ considerably. We identify two main relevant approaches to model item novelty, based on discovery and distance respectively, which we describe next. The framework is nonetheless open to the modular integration of alternative models.

4.3.1 Popularity-Based Item Novelty

In a generic sense, item novelty can be defined as the difference between an item and "what has been observed" in some context. The notion of item discovery introduced in the previous section enables a formulation of this principle as the probability that an item was not observed before:

$$nov(i|\theta) = 1 - p(seen|i,\theta) \quad (4.3)$$

The contextual variable θ here represents any element on which item discovery may depend, or relative to which we may want to particularize novelty. This might include e.g. a specific user, a group of users, vertical domains, time intervals, sources of item discovery –such as searching, browsing, past or alternative recommendations, friends, advertisements, etc. The specific instantiation of θ we develop here consists of the observed interactions between users and items, available to the system under evaluation. We will nonetheless briefly discuss in section 4.6.3 other interesting metrics that result when considering alternative contexts.

In general terms, $p(seen|i,\theta)$ reflects a factor of item popularity, whereby high novelty values correspond to long-tail items few users have interacted with, and low novelty values correspond to popular head items. If we wish to emphasize highly novel items, we may also consider the log of the inverse popularity:

$$nov(i|\theta) = -\log_2 p(seen|i,\theta) \quad (4.4)$$

Alternatively, one may also consider the Bayesian inversion of the discovery distribution, $p(i|seen,\theta)$, which provides a relative measure of how likely items are to be seen with respect to each other. This leads to an interesting formulation of item novelty:

$$nov(i|\theta) = -\log_2 p(i|seen,\theta) \quad (4.5)$$

This corresponds to the notion of self-information or surprisal $I(i)$, commonly used in Information Theory to measure novelty as the amount of information the observation of i conveys (Zhou, Kuscsik, Liu, Medo, Wakeling, & Zhang, 2010). Interestingly, this distribution –to which we will refer as *free discovery*– can be directly connected to the previous one –which we will term *forced discovery*. Assuming items are sampled uniformly in the absence of discovery conditions –i.e. we assume a uniform $p(i|\theta)$–, it can be seen that $p(i|seen,\theta) = p(seen|i,\theta)/\sum_{j\in\mathcal{I}} p(seen|j,\theta)$. The free and forced discovery models are therefore equivalent except for a normalizing constant $\sum_{j\in\mathcal{I}} p(seen|j,\theta)$ that depends only on θ. In our experiments we have found that this constant does not introduce a significant difference in the resulting metrics, which suggests that both models –free and forced discovery– could be used indistinctly.

4.3.2 Distance-Based Item Novelty

The novelty model scheme defined in the previous section considers how different an item is from past experience in terms of strict Boolean identity: an item is new if it is absent from past experience ($seen = 0$) and not new otherwise ($seen = 1$). There are reasons however to consider relaxed versions of the Boolean view: the knowledge available to the system about what users have seen is partial, and therefore an item might be familiar to a user even if no interaction between them has been observed in the system. Furthermore, even when a user sees an item for the first time, the resulting information gain –the effective novelty– ranges in practice over a gradual rather than binary scale (consider for instance the novelty involved in discovering the movie "Rocky V").

As an alternative to the popularity-based view, we consider a similarity-based model where item novelty is defined by a distance function between the item and a context of experience. If the context can be represented as a set of items, for which we will intentionally reuse the symbol θ, we can formulate this as the *expected* or *minimum distance* between the item and the set:

$$nov(i|\theta) = \sum_{j \in \theta} p(j|choose, \theta, i) d(i,j)$$

$$\text{or} \quad nov(i|\theta) = \min_{j \in \theta} d(i,j)$$

where $p(j|choose, \theta, i)$ is the probability that the user chooses item j in the context θ, when she has already chosen i. The distance measure d can be defined e.g. as the complement $d(i,j) = 1 - sim(i,j)$ of some similarity measure (cosine-based, Pearson correlation, etc., normalized to [0,1]) in terms of the item features –content-based view– or their user interaction patterns –collaborative view. Assuming a uniform $p(j|\theta)$, it can be seen that:

$$nov(i|\theta) = \frac{\sum_{j \in \theta} p(choose|j, \theta, i) d(i,j)}{\sum_{j \in \theta} p(choose|j, \theta, i)} \tag{4.6}$$

where the denominator acts as a normalizing constant for θ. The forced choice probability is easier to compute than its free counterpart, as we shall see, and has a somewhat clearer interpretation: $p(choose|j, \theta, i)$ weights the sum in a way that the distance $d(i,j)$ is only counted if the user actually cared about j. This term plays a similar role as in equation 4.2, and can be developed as a browsing model –see next section–, or simplified to $p(choose|j, \theta, i) \sim 1$, in which case $nov(i|\theta)$ just becomes an average distance.

In the context of distance-based novelty, we find two useful instantiations of the θ reference set: a) the set of items a user has interacted with –i.e. the items in her profile–, and b) the set R of recommended items itself. In the first case, we get a user-relative novelty version of equation 4.6, and in the second case, we get the basis for a generalization of intra-list diversity, as we will show. It is possible to explore other possibilities for θ, such as groups of user profiles, browsed items over an interactive session, items recommended in the past or by alternative systems, etc., which we leave as future work.

4.4 Browsing Model

The browsing component of the metric scheme, as introduced in equation 4.2, is based on a distribution $p(choose|i, u, R)$ which we may model in terms of the user behavior in its interaction with a list of recommended items. There are many ways to model this behavior. Our approach takes inspiration in related work on user click models in information retrieval systems (Carterette, SIGIR, 2011, Clarke, et al., 2008, Hu, Zhang, Chen, & Wang, 2011, Moffat & Zobel, 2008 and Radlinski, Kleinberg, & Joachims, 2008), but any other alternative modeling approach could be plugged into our framework.

Our model goes as follows. First, we consider the target user will use all recommended items which she effectively gets to see *and* finds relevant for her taste. We had already formulated this view in equation 4.1, which in the current context becomes:

$$p(choose|i, u, R) \sim p(seen|i, u, R) p(rel|i, u)$$

where we assume the relevance of an item is independent from the recommendation in which it is delivered. The $p(rel|i, u)$ component introduces relevance in the definition of the metric: the novelty of a recommended item will be taken into account only as much as the item is likely to be relevant for the target user.

The $p(seen|i,u,R)$ component represents the probability that the target user will actually see the item i when she is browsing the ranked list R. This component allows for the introduction of a rank discount by having $p(seen|i,u,R)$ reflect the fact that the lower an item is ranked in R, 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 C. , Craswell, Soboroff, & Ashkan, 2011) 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 $p(cont|k,u,R)$ is the probability that user u decides to continue browsing the next item at position $k+1$. With this scheme we have, by recursion:

$$\begin{aligned} p(seen|i_k,u,R) &= p(seen|i_{k-1},u,R)p(cont|k-1,u,R) \\ &= \prod_{l=1}^{k-1} p(cont|l,u,R) \end{aligned} \tag{4.7}$$

Now there are several ways –of varying complexity– in which $p(cont|l,u,R)$ can be modeled. A simple one is to consider a constant $p(cont|l,u,R)=p_0$, whereby we get an exponential discount $p(seen|i_k,u,R)=p_0^{k-1}$. This is the approach taken in the RBP search performance metric (Moffat & 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 $p(seen|i_k,u,R)=\prod_{l=1}^{k-1}\big(1-p(rel|i_l,u)\big)$, similar to the ERR metric (Chapelle, Metzler, Zhang, & Grinspan, 2009), or the models in (Radlinski, Kleinberg, & Joachims, 2008). We might consider more complex and general models, such as:

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

similar to (Clarke, et al., 2008), or $p(cont|l,u,R)=p(cont|rel)p(rel|i_l,u)+p(cont|\neg rel)\big(1-p(rel|i_l,u)\big)$, and so forth. In general, we may use any decreasing rank discount function $p(seen|i_k,u,R)=disc(k)$ we deem suitable, even heuristic ones, such as a logarithmic discount as in nDCG, a Zipfian discount, etc., or even no discount by $disc(k)=1$, as if the user always browsed the whole list. Putting all this together, equation 4.2 can be rewritten as a configurable rank-sensitive, relevance aware metric scheme:

$$m(R|\theta) = C\sum_{i_k\in R} disc(k)p(rel|i_k,u)nov(i_k|\theta) \tag{4.8}$$

We are now in a position to define the normalizing constant C, which is intended to stabilize the metric against unwanted biases. Two normalization approaches are commonly considered in information retrieval metrics, which define $1/C$ respectively as: a) the maximum metric value obtainable by an ideal recommendation ranking, e.g. as in nDCG and α-nDCG (Clarke, et al., 2008), or b) the expected browsing depth, as in RBP (Moffat & Zobel, 2008) and discussed in (Clarke C. , Craswell, Soboroff, & Ashkan, 2011). Computing the ideal ranking is metric-specific and often costly,

sometimes even NP-hard, though it can be approximated by greedy approaches (Clarke, et al., 2008). The expected browsing depth is more straightforward to compute:

$$\frac{1}{C} = \sum_{i_k \in R} k \cdot p(seen|i_k, u)\big(1 - p(cont|i_k, u)\big)$$

$$= \sum_{i_k \in R} k\big(disc(k) - disc(k+1)\big) = \sum_{i_k \in R} disc(k)$$

where we define $disc(k) = 0$ if $k > |R|$ (i.e. $p(seen|i, R) \sim 0$ if $i \notin R$). It can be seen that with no rank discount ($disc(k) = 1$) we have $C = 1/|R|$ (average relevance-weighted item novelty).

In order to make this scheme fully implementable, we need to provide practical methods to estimate the primary models –discovery and relevance– upon which we have built the framework, based on observed data. We do this in the next section.

4.5 Estimation of Ground Models

4.5.1 Item Discovery

The estimation of the discovery model depends on our definition of θ and the type of available data. If we take θ as the set of observed interactions between users and items in the system, and the data consists of user ratings for items represented as a functional relation $\theta \equiv r : \mathcal{U} \times \mathcal{I} \to \mathcal{V}$, we may take a maximum likelihood model estimate by:

$$p(seen|i, r) \sim \frac{|\mathbf{i}|}{|\mathcal{U}|} = \frac{|\{u \in \mathcal{U} | r(u, i) \neq \emptyset\}|}{|\mathcal{U}|} \tag{4.9}$$

where $\mathbf{i}$ denotes the set of users who have rated i, and $r(u, i) \neq \emptyset$ means the rating of u for i is known. If the available data consists of implicit preference observations in the form of a set $\theta \equiv \mathcal{L}$ of user/item/timestamp records, the estimate would be:

$$p(seen|i, \mathcal{L}) \sim \frac{|\mathbf{i}|}{|\mathcal{U}|} = \frac{|\{u \in \mathcal{U} | \exists t \in \mathcal{T} : (u, i, t) \in \mathcal{L}\}|}{|\mathcal{U}|} \tag{4.10}$$

$\mathcal{T}$ being the timestamp data type. Note that with these estimates, item novelty in equation 4.4 becomes the *inverse user frequency* IUF. The free novelty model can also be estimated over ratings or implicit data, respectively, as:

$$p(i|seen, r) \sim \frac{|\mathbf{i}|}{\sum_{j \in \mathcal{I}} |\mathbf{j}|} = \frac{|\{u \in \mathcal{U} | r(u, i) \neq \emptyset\}|}{|\{(u, j) \in \mathcal{U} \times \mathcal{I} | r(u, j) \neq \emptyset\}|} \tag{4.11}$$

$$p(i|seen, \mathcal{L}) \sim \frac{|\mathbf{i}|}{\sum_{j \in \mathcal{I}} |\mathbf{j}|} = \frac{|\{u \in \mathcal{U} | \exists t \in \mathcal{T} : (u, i, t) \in \mathcal{L}\}|}{|\{(u, j) \in \mathcal{U} \times \mathcal{I} | \exists t \in \mathcal{T} : (u, i, t) \in \mathcal{L}\}|} \tag{4.12}$$

With the rating-based estimate (equation 4.11), equation 4.5 becomes the so-called *inverse collection frequency* ICF.

4.5.2 Item Relevance

Relevance in the context of recommendation is a user-specific notion which can be equated to the interest of users for items. How relevance can be modeled depends again on the nature of available observations. If the available input consists of explicit user ratings, the probability of items being liked can be modeled by a heuristic mapping

between rating values and probability of relevance. For instance, drawing from the ERR metric scheme (Chapelle, Metzler, Zhang, & Grinspan, 2009):

$$p(rel|i,u) \sim \frac{2^{g(u,i)}-1}{2^{g_{max}}} \tag{4.13}$$

where g is a utility function to be derived from ratings, e.g. $g(u,i) = \max(0, r(u,i) - \tau)$, where τ represents the "indifference" rating value, as described by Breese, Heckerman, & Kadie (1998). In our experiments we try a slight variation with respect to (Chapelle, Metzler, Zhang, & Grinspan, 2009): we do not subtract 1 in the numerator in order to avoid a drastic loss of novelty signal by overfitting to zero the probability of unobserved relevance.

For usage logs, a correspondence can be fairly established between item usage counts and user interest, which we account for in two steps. First, we normalize the observed item access frequencies of each user to a common rating scale $[0, n]$, as proposed in (Celma & Herrera, 2008). Namely, $r(u,i) \longleftarrow n \cdot F\left(frec_{u,i}\right)$, where $frec_{u,i}$ is the number of times u has accessed i, and $F\left(frec_{u,i}\right) \sim \left|\{j \in \mathbf{u} | f_{u,j} \leq f_{u,i}\}\right| / |\mathbf{u}|$ is the cumulative distribution function of $frec_{u,i}$ over the set of items in the profile of u – denoted as $\mathbf{u}$. Then we apply to these ratings the same mapping as before (equation 4.13), this time with $\tau = 0$ –assuming that accessing an item, however infrequently, does not in general reflect a negative preference.

4.6 Recommendation Novelty and Diversity Metrics

4.6.1 Novelty

By plugging the popularity-based item novelty models (section 4.3.1) in the general metric scheme (equation 4.8), we get discovery-based recommendation novelty metrics. For instance, taking equation 4.3, we get:

$$\text{EPC} = C \sum_{i_k \in R} disc(k) p(rel|i_k,u)\left(1 - p(seen|i_k)\right) \tag{4.14}$$

which we label as expected popularity complement (EPC). Equations 4.4 and 4.5 similarly lead to alternative formulations, to which we shall refer as expected inverse popularity (EIP), and expected free discovery (EFD), respectively. All three metrics provide a measure of the ability of a system to recommend relevant long-tail items. EPC can be read as the expected number of seen relevant recommended items not previously seen. EIP and EFD can be read as the expected IUF and ICF of (relevant and seen) recommended items, respectively. Note that if we ignore rank and relevance, then $\text{EFD} = -\frac{1}{|R|}\sum_{i \in R} \log_2 p(i|seen)$, the mean self-information (MSI) of the recommended items, a metric reported in (Zhou, Kuscsik, Liu, Medo, Wakeling, & Zhang, 2010).

If we take a distance-based novelty model (equation 4.6) relative to the set of items the target user has interacted with $\theta \equiv \mathbf{u}$ –i.e. the items in her profile– we get an alternative novelty measure consisting of the expected distance between the recommended items and the items in the user profile, which we label as the expected profile distance (EPD):

$$\text{EPD} = C' \sum_{i_k \in R, j \in \mathbf{u}} disc(k) p(rel|i_k,u) p(rel|j,u) d(i_k,j) \tag{4.15}$$

where $C' = C/\sum_{j \in \mathbf{u}} p(rel|j,u)$. In this case, each term in the summation is doubly weighted by the relevance of the involved item pair, and only once by the rank distance function. This is because we assume $p(seen|i) = 1$ for items in the user profile. The metric provides a user-relative measure of novelty which, as far as we are aware of, has not been reported in the literature.

4.6.2 Diversity

In the distance-based model, if we take $\theta \equiv R$, we get a measure of recommendation diversity:

$$\text{EILD} = \sum_{\substack{i_k, i_l \in R \\ k \neq l}} C_k disc(k) disc(l|k) p(rel|i_k, u) p(rel|i_l, u) d(i_k, i_l) \tag{4.16}$$

where $disc(l|k) = disc(\max(1, l-k))$ reflects a relative rank discount for an item at position l knowing that position k has been reached. This general form provides a doubly rank-sensitive and rank-aware expected intra-list diversity metric. In this case the normalizing constant is $C_k = C/\sum_{i_l \in R - \{i_k\}} disc(l|k) p(rel|i_l, u)$. If we remove the rank discount and relevance weighting, the metric reduces to:

$$div(R|u) = \frac{2}{|R|(|R|-1)} \sum_{i_k \in R, l < k} d(i_k, i_l) = \text{ILD}$$

Equation 4.16 thus generalizes the average intra-list distance (ILD) (Zhang & Hurley, RecSys, 2008 and Ziegler, McNee, Konstan, & Lausen, 2005) with the introduction of rank-sensitivity and relevance.

4.6.3 Further Unification

By explicitly modeling novelty as a relative notion, the proposed framework has a strong unifying potential of further novelty and diversity conceptions. In other to illustrate this, let us consider the notion of temporal diversity proposed in (Lathia, Hailes, Capra, & Amatriain, 2010), which we will refer to as self-system diversity (SSD). It is defined as the ratio of recommended items that were not included in a previous recommendation:

$$\text{SSD}(R|u) = \frac{|R \setminus R_{t-1}|}{|R|} \tag{4.17}$$

R_{t-1} being the last recommendation delivered by the system for u before R. This notion can be described in our framework in terms of a discovery model where the source of discovery is the last recommendation, as follows. Taking $\theta \equiv \langle u, R_{t-1} \rangle$ as the context of discovery, we get $p(seen|i,\theta) = p(seen|i,u,R_{t-1}) = disc(i|R_{t-1})$, where the latter represents the discount that corresponds to the position of i in R_{t-1} (0 if $i \notin R_{t-1}$). Thus, the novelty of an item is defined by a browsing model over the last recommendation. Plugging this into the general metric scheme gives:

$$div(R|u) = \text{ESSD} = C \sum_{i_k \in R} disc(k) p(rel|i_k, u)\big(1 - disc(i_k|R_{t-1})\big)$$

If we ignore rank and relevance in R, and rank in R_{t-1} –that is, we take $p(seen|i,u,R_{t-1}) \sim 1_{R_{t-1}}(i)$– it can be seen that we get the original SSD expression in equation 4.17. Thus our framework provides again a formalization and generalization of the metric with the possibility to easily introduce rank and relevance.

This scheme can be similarly applied to other novelty and diversity metrics, such as temporal novelty as defined in (Lathia, Hailes, Capra, & Amatriain, 2010), inter-system novelty (novelty of recommended items with respect to recommendations that alternative systems may procure), or inter-user diversity (with respect to the recommendations other users are getting) as defined in (Bellogín, Cantador, & Castells, 2010). Table 2 summarizes some of the metrics that can be unified in our framework by different instantiations of θ in the item novelty scheme.

Metric scheme	Context θ	User perspective	Generalizes
Long tail (popularity)	Ratings r or frequencies $\mathcal{L}$	Novelty	Mean self-information (Zhou, Kuscsik, Liu, Medo, Wakeling, & Zhang, 2010)
Distance-based	Target user u	Novelty	-
	Recommendation R	Diversity	Intra-list diversity (Zhang & Hurley, RecSys, 2008) (Ziegler, McNee, Konstan, & Lausen, 2005)
Alternative discovery sources	Last recommendation $\langle u, R_{t-1} \rangle$	Novelty	Self-system diversity (Lathia, Hailes, Capra, & Amatriain, 2010)
	All previous recommendations $\langle u, A_{t-1} \rangle$	Novelty	Self-system novelty (Lathia, Hailes, Capra, & Amatriain, 2010)
	Recommendations by other systems $\langle u, \mathcal{S} \rangle$	Novelty	Inter-system novelty (Bellogín, Cantador, & Castells, 2010)
	Recommendations to other users $\langle \mathcal{U}, s \rangle$	Novelty	Inter-user diversity (Bellogín, Cantador, & Castells, 2010)

Table 2. Unification of state of art novelty and diversity metrics in the proposed metric framework.

4.7 An Example

In order to illustrate the effects of the proposed metrics, and in particular the rank discount and relevance weighing, we show here the computation of some variants over a small artificial example. We select the EPC metric scheme (equation 4.14), which for illustrative purposes is representative of similar effects in the other metrics.

Assume we have a system with 1,000 users, and a target user u with 8 items in her profile. For simplicity, assume the rating scale is binary {0,1}, with indifference value $\tau = 0$. Assume we have two systems which deliver recommendations R_1 and R_2 to u respectively, with the content shown in Table 3. In the example we just show the known rating value $r(u, i)$ of each item by the target user (i.e. relevance), and the popularity of the items in terms of the number of users who have rated each. It is easy to see that both recommendations do equally well in terms of returned relevant items, but R_2 does a better job at ranking long-tail items (with few ratings) by the top of the list.

	R_1		R_2	
Position	$r(u,i)$	# raters	$r(u,i)$	# raters
1	1	1000	1	10
2	1	1000	1	10
3	1	500	1	10
4	1	500	1	500
5	1	10	1	500
6	1	10	1	1000
7	1	10	1	1000
8	0	10	0	1000
9	0	10	0	10
10	0	10	0	10

Table 3. An illustrative example.

Based on equations 4.9 and 4.13 for discovery and relevance model estimation respectively, and using a logarithmic rank discount $disc(k) = 1/\log_2(k+1)$, we get the metric values shown in Table 4. The best result is bolded for each metric. According to EPC ignoring relevance and rank, R_1 performs better than R_2, because it includes an equal number of relevant items, but a more novel, long-tail item in position 8 (with 10 vs. 1000 ratings). EPC_{rel} does not count this difference because the item at that position is not relevant, whereby both lists get the same metric value. Considering rank but not relevance, EPC_{rank} detects that R_1 does a poor job at ranking the novel items in the list compared to R_2, even if the novel item at position 8 is appreciated by the metric (which does not care that the item is non-relevant). Combining both rank and relevance, R_2 scores best, by the highest difference of all metrics. If we agree that R_2 is objectively better than R_1, $EPC_{rank,rel}$ is the metric that best discriminates this fact.

To compensate for the lack of relevance awareness of diversity metrics, prior work has used complementary accuracy measures. To further illustrate the utility of a configuration integrating rank and relevance-awareness in a single metric, as opposed to the combination of two separate measures, we show in the last row of the table one such combination: the harmonic mean of nDCG (pure accuracy, rank aware) and EPC (pure novelty). This combined metric prefers R_1 to R_2 because it has one more novel item at position 8. But the metric fails to realize that this item is not relevant, and furthermore it disregards the fact that all the novel items aside this one are sorted fairly worse in R_1 than in R_2. In contrast, $EPC_{rank,rel}$ does not suffer from these shortcomings.

	$disc(k)$	$p(rel\|i,u)$	R_1	R_2
nDCG	-	-	**0.9202**	**0.9202**
EPC	1	1	**0.6940**	0.5950
EPC_{rank}	$1/\log_2(k+1)$		0.5343	**0.6829**
EPC_{rel}	1	$\frac{2^{g(u,i)}-1}{2^{g_{max}}}$	**0.3970**	**0.3970**
$EPC_{rank,rel}$	$1/\log_2(k+1)$		0.3370	**0.5543**
H (nDCG, EPC)	1	1	**0.7913**	0.7227

Table 4. Resulting values of different metrics for the two example recommendations, combining different rank and relevance configurations in the EPC novelty metric.

4.8 Experimental Results

We have tested our framework in different metric configurations on two datasets –explicit and implicit data– with several baseline recommenders and diversification methods. On the one hand, we have used the MovieLens 1M dataset, which includes one million ratings by 6,040 users for 3,900 items. For an implicit preference dataset, we have used an extract from Last.fm provided by Celma & Herrera (2008), including the full listening history of 992 users till May 2009. The data involves 176,948 artists and a total of 19,150,868 user accesses to music tracks. For the computation of the proposed metrics, the data are split into training and test sets. In MovieLens we use the five 80-20% rating splits provided in the dataset distribution, providing for 5-fold cross-validation. In the Last.fm dataset, we apply a temporal split leaving 80% of scrobblings in the "past" for training, and the 20% most recent for testing.

	CB	MF	UB	AVG	RND
MovieLens1M	0.1113	0.2136	0.1463	0.1497	0.0332
Last.fm	-	0.3081	0.5797	0.0392	0.0107

Table 5. Accuracy of the tested baselines, measured in nDCG@50 over the two datasets.

We run three representative state of the art recommender system algorithms on the two datasets, namely, a user-based kNN recommender with 100 neighbors (UB), a matrix factorization algorithm (Koren, Bell, & Volinksy, 2009) with 50 latent factors (MF), and a content-based algorithm (CB). The latter is only tested on MovieLens using movie genres, as the Last.fm dataset does not include content features to support a CB recommender. For further reference, we test two additional probe baselines: average rating (AVG), and random recommendation (RND). The recommenders are run on Last.fm by mapping access frequencies to ratings as proposed in (Celma & Herrera, 2008), taking artists as items. In order to give a reference on the behavior of the baselines in terms of accuracy, we show their nDCG@50 in Table 5.

The discovery models (equations 4.3-4.5) are built on training data –since they do not involve target users– and the relevance models (equation 4.13) on test data. The estimation of the discovery models is based on equations 4.9 and 4.11 for MovieLens (explicit ratings) and equations 4.10 and 4.12 for Last.fm (item access log). The browsing models build exclusively on test data (for relevance, equation 4.13) and recommenders' output (for recommendation discovery distribution, equation 4.7). The distance-based metrics compare items in terms of their genres in MovieLens, and their test ratings in Last.fm, as the complement of the Jaccard and Pearson similarities (shifted to [0,1]), respectively. We measure all metrics at a top 50 ranking cutoff.

4.8.1 Pure and Relevance-Aware Metrics

Figure 11 shows how the tested recommenders compare on different metrics, namely EPC, EPD, and EILD (equations 4.14, 4.15, 4.16). We omit EIP (log of inverse popularity), and EFD (free discovery model) as they yield equivalent measurements to EPC –aside a matter of scale– in terms of the relative comparison of recommenders in all configurations. We first focus on the relevance-unaware metric versions (top two graphics in the figure). A first interesting observation is that CB is better than the CF recommenders in popularity-based novelty (confirming findings in Celma & Herrera,

2008), but is worse at diversity and user-specific novelty. This is what one would expect: CB concentrates recommendations around the users' profile, hereby scoring low on EPD. Being similar to the profile, recommended items are also similar among themselves, which explains the low EILD. UB and MF avoid such shortcomings, but they tend to concentrate recommendations on items with enough available ratings to infer recommendations. Hence they have a bias towards popular items –penalized by the popularity-based metrics– which CB does not suffer from (this is related to the well-known suitability of CB for cold-start items). AVG does not show any particular trend, as it is mostly independent from popularity and the other signals the metrics are sensitive to. Note that in AVG we apply a linear rating penalization on items with less than five raters, to avoid single-rater favorites (as low-confidence averages) to swamp the top of recommendations –in which case AVG would score much higher on novelty. Finally, random recommendation gets the highest values in all relevance-unaware metrics (except for some near ties on MovieLens), illustrating the fact that pure novelty and diversity metrics alone are not enough –note to this respect that such configurations of EILD and EPC (insensitive to rank and relevance) correspond to state of the art metrics (Zhang & Hurley, RecSys, 2008, Zhou, Kuscsik, Liu, Medo, Wakeling, & Zhang, 2010 and Ziegler, McNee, Konstan, & Lausen, 2005).

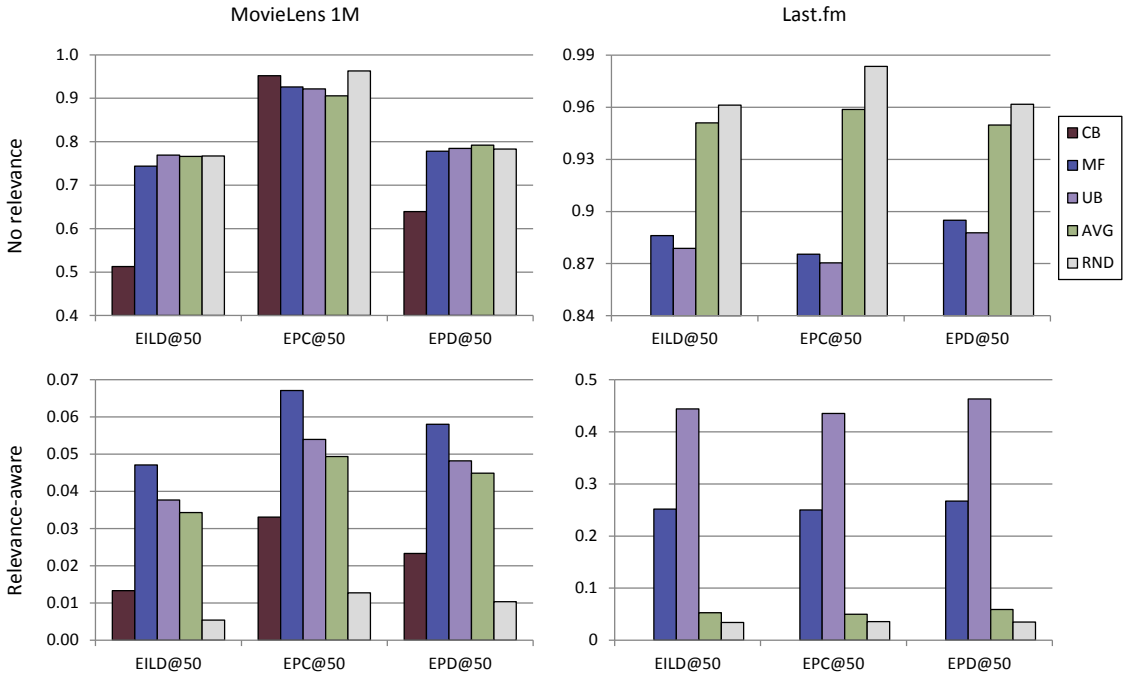

Figure 11. Novelty and diversity metrics are shown on four baselines (content-based, matrix factorization, user-based kNN, average, and random) over MovieLens 1M –two graphics on the left– and Last.fm –right. The top two graphics display metrics that ignore relevance, whereas the bottom ones are relevance-aware. All the metrics in the figure are rank-insensitive.

The two bottom graphics in Figure 11 show the relevance-aware variant of the metrics. With this configuration MF takes the lead on MovieLens data. It was very similar to UB on pure novelty, but it beats UB on relevance (see Table 5), and has a good trade-off between novelty and relevance compared to the other recommenders. The reverse situation occurs on Last.fm, where UB has higher accuracy than MF. Random gets a drastic drop in both cases for its lack of accuracy –to which respect this metric variant thus behaves better than the pure novelty and diversity metrics. CB gets a

noticeable decrease as well, for a similar (though not as extreme) reason. The lesser quality of AVG recommendations –hence their lower actual ratio of useful diversity– is also evidenced by relevance awareness, particularly in Last.fm.

4.8.2 Rank Sensitiveness

Rank-aware metric configurations should not discriminate the baselines much further than this, since none of the recommenders target novelty, and whatever amount they get is by unsought reasons –their share of novelty is randomly ordered. In order to test rank sensitivity, we set up three diversification strategies that do optimize for novelty and diversity. The diversifiers re-rank the top n recommended items (n = 500 in our experiment) returned by a baseline recommender, by greedily optimizing an objective function. Specifically, we adapt a) the diversification strategy proposed in (Ziegler, McNee, Konstan, & Lausen, 2005), which we term Maximal Marginal Relevance (MMR) for its connection to the approach described in (Carbonell & Goldstein, 1998), where the objective function is a trade-off of accuracy and diversity –namely, a linear combination (we take equal weights $\lambda = 0.5$) of the baseline rating prediction (accuracy) and the average dissimilarity to the items above each position (diversity); b) a variant of the latter, which we call novelty-based greedy diversification (NGD), where a function targeting unpopularity (IUF as defined by equation 4.4) is used in place of the dissimilarity component; and c) an adaptation of the IA-Select algorithm (Agrawal, Gollapudi, Halverson, & Ieong, 2009), originally devised for search diversification and adapted in chapter 3. Additionally, we include a random re-ranking.

Table 6 shows the results on diversifying the MF baseline, confirming consistent trends with the sought metric properties. We may observe, first, that without relevance, few diversifiers beat the random re-ranking, although some do –e.g. NGD on EPC, consistently with its quite specific optimization target. However, with relevance, random is always worst, except for NGD on MovieLens: this diversifier promotes unpopular items, which tend to score low on overall relevance –still, with rank discount NGD also beats the random approach. IA-Select seems to be the best diversifier in terms of the trade-off between relevance and diversity. Its results particularly stand out on Last.fm with relevance, even better with rank discount, and best of all on EILD, since this algorithm specifically targets diversity, above novelty. It can also be seen that the baseline is less easy to beat in the relevance-aware metrics, although some diversifiers manage to do so, most-notably IA-Select.

We may also observe that the rank discount (we test $disc(k) = 0.85^{k-1}$ based on Moffat & Zobel, 2008) changes the sign of comparison in several cases. To point out a few: without relevance, this occurs for IA-Select vs. MMR on EPC and vs. the baseline on EPD, on Last.fm, or IA-Select vs. the baseline on EPC on MovieLens. On Last.fm with relevance, NGD switches from underperforming to overperforming the baseline and MMR on all three metrics. The difference in IA-Select captured by adding rank to EILD with relevance in Last.fm is particularly noteworthy. All these examples show how the rank sensitivity uncovers improvements that would otherwise go unnoticed.

4.9 Conclusion

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. As future work, we plan to complement our off-line experiments with on-line tests where the different metric configurations are contrasted to actual user feedback on the recommendation quality and utility aspects we seek to measure.

			EPC@50		EPD@50		EILD@50	
		$disc(k)$	1	0.85^{k-1}	1	0.85^{k-1}	1	0.85^{k-1}
MovieLens1M	No relevance	MF	0.9124	0.8876	0.7632	0.7466	0.7164	0.6191
		IA-Select	*0.9045*	0.8886	**0.8080**	0.7577	**0.8289**	**0.7483**
		MMR	*0.9063*	*0.8769*	*0.7605*	*0.7428*	0.7191	0.6247
		NGD	**0.9851**	**0.9795**	**0.7725**	0.7551	*0.6563*	*0.5430*
		Random	0.9525	0.9527	0.7699	0.7699	0.7283	0.6719
	Relevance	MF	**0.0671**	**0.1043**	**0.0580**	**0.0944**	**0.0471**	**0.0551**
		IA-Select	**0.0705**	**0.1161**	**0.0639**	**0.1032**	**0.0537**	**0.0648**
		MMR	**0.0719**	**0.1131**	**0.0620**	**0.1020**	**0.0510**	**0.0610**
		NGD	*0.0155*	***0.0223***	*0.0128*	***0.0200***	*0.0067*	*0.0017*
		Random	*0.0222*	*0.0218*	*0.0182*	*0.0179*	*0.0117*	*0.0058*
Last.fm	No relevance	MF	0.8754	0.8481	0.8949	0.8895	0.8862	0.7954
		IA-Select	0.8840	0.9089	*0.8912*	(0.8909)	(0.8878)	0.8274
		MMR	0.9068	0.8903	0.9133	0.9107	0.9166	0.8398
		NGD	**0.9722**	**0.9571**	**0.9423**	**0.9398**	**0.9485**	**0.8784**
		Random	0.9359	0.9357	0.9278	0.9279	0.9318	0.8619
	Relevance	MF	**0.2501**	**0.2115**	**0.2671**	**0.2587**	**0.2518**	**0.1900**
		IA-Select	**0.3343**	**0.4752**	**0.3462**	**0.3994**	**0.3343**	**0.4154**
		MMR	***0.2351***	***0.1936***	***0.2439***	***0.2340***	***0.2360***	***0.1759***
		NGD	***0.2286***	**0.3077**	***0.2212***	(**0.2593**)	***0.2165***	**0.2656**
		Random	*0.1362*	*0.1368*	*0.1407*	*0.1405*	*0.1342*	*0.1113*

Table 6. Results on EPC, EPD, EILD on different diversifications of the MF baseline recommender, with all relevance and rank discount combinations. For the rank-sensitive variants an exponential discount is used as in (Moffat & Zobel, 2008)**, with power base 0.85. Values better than random are in bold, values below the baseline in italics, and the best recommendation for each metric is underlined. All differences with respect to random and baseline are statistically significant (Wilcoxon p < 0.001) except when in parenthesis (respect to the MF baseline).**

5. Conclusions

5.1 Summary and Contributions

Novelty and diversity for RS and IR have received increasing interest in the last years. Both properties are essential in RS for real-word scenarios and applications –such as online commerce–, where the aggregated relevance of individual items or document does not necessarily guarantee an optimal or even satisfactory user experience. Novelty and discovery are directly linked with avoiding the monotony and obviousness of recommendations, thus improving the capacity of discovery and broadening and enriching the user experience.

Such concerns are relatively recent in IR and RS, many open questions remain and motivate further research. For one, the RS and IR communities have addressed novelty and diversity quite differently, which brings about the opportunity to investigate connections, equivalences and differences.

Our work focuses in the enhancement and evaluation of novelty and diversity in RS, specifically:

- Concerning the enhancement, we have proposed an adaptation of diversification methods of IR to RS, such as the MMR and IA-Select algorithms, by defining the concept of aspect space to represent the variety of interests of the user as equivalence to query intents. We extract aspect spaces from item feature data and latent information between users and item.
- Regarding the evaluation, we also used aspect spaces to adapt intent-aware metrics to RS, such as ERR-IA, nDCG-IA and α-nDCG, and provided a unified metric framework that includes some of the state of the art metrics for novelty and diversity in RS, such as ILS, MSI, and others, and allows their extension to consider further important properties of recommendation lists such as the rank and relevance of recommended items. We derived and formalized further metrics based on inverse popularity, such as EPC, EIP and EFD, based on the distance with the user profile as EPD and based on the distance between the elements of a recommendation list, such as EILD.
- Experiments conducted on datasets from MovieLens and Last.fm provided empiric evidence of the effectiveness of our IR diversification approach and allowed for observation and analysis of the characteristics of the different metrics derived from the framework.

Our contributions provide new ways of stating, formalizing and addressing the problems of novelty and diversity in RS. On one side our adaptation of IR diversity techniques to RS allows the advances made on the first to be applied on the latter. On the other side, the metric framework for novelty and diversity provides a platform for analysis of common components of metrics based on simple properties of recommendation lists. We thus aim to elaborate a more comprehensive perspective which focuses on user satisfaction rather than just individual accumulated relevance or error minimization, as has been traditionally the case in the field.

5.2 Discussion and Future Work

Several topics addressed here motivate further analysis and research. We point here towards some of our ongoing research directions and improvements in progress.

5.2.1 Explicit Aspect Spaces Extraction

The goodness of item features for diversification is a factor that has not been explicitly examined. Consider the case of the experiments with the MMR algorithm in chapter 3. Since we used movie genres as features as aspects, it is very likely that the possible distance values defined by set intersection reduce to a very small set, since the set of genres is relatively small (18) and movies usually have few of them (1.72 in average). This would not be the case of implicit aspect spaces, where a higher-dimensional vector space with real valued components is created. Other aspects, such as the distribution of features over items, can be determinant to the effectiveness of diversification algorithms. Therefore we doubt that every item features set, though representative, could be flexible enough to our diversification methods.

One possible solution to the previous problems would be the use more than one feature type (genre, language, country, location, etc.) to balance or compensate weaknesses of separated features. Some initial work towards the determination of goodness of item features has been done in (Vargas, Castells, & Vallet, 2012).

5.2.2 Implicit Aspect Spaces Extraction

Continuing with the extraction of aspect spaces, we made a direct adaptation of the implicit factors of matrix factorization algorithms as item features and proceeded like the case of explicit features. While this approach works well, it is possible to refine it, e.g. beyond a binary approach. Since vector dimensions may have different magnitudes (in terms of variance) the selection of dimensions for an item may be biased to factors with high magnitude but little relative significance to the item. A normalization step might provide enhanced results on this point. The binarization itself may in fact no longer be required after a normalization of the vector space dimensions. A probabilistic model that would transform smoothly from vector components q_i to probabilities $p(f|i)$ might be applied instead, and it should also deal with negative values of the vector components. Finally, in our proposal we estimated $p(f|u)$ as we did in the case of item features, where there is no direct information between features and users, but this is not the case. In fact, each user has a vector p_u that represents her in the vector space, so using the same procedure as for items it would be possible to derive $p(f|u)$ directly.

Additionally, we are currently working in the application of other sources of latent analysis techniques. Specifically, we are considering the probabilistic Latent Semantic Analysis of (Hofmann, 2004) and Latent Dirichlet Allocation of (Blei, Ng, & Jordan, 2003). Both have a probabilistic formulation and have been used in RS for building powerful recommenders, so we think that they would be suitable for the definition of new implicit aspect spaces for diversification.

5.2.3 Diversification methods

We would like to extend our adaptation or IR diversity to other algorithms. As an inmediate work, we are currently adapting the xQuAD algorithm of Santos, Macdonald, & Ounis (WWW, 2010).

We are also interested in the revision of some fundamental models that are implicit in IR diversification methods such as IA-Select and xQuAD where, parting from a generative model, the probability of choice or selection is modeled as a disjoint probability over the set of items (in IA-Select $V(d|c,q) = p(d|c,q)$) instead of considering an alternative based on multiple selection and relevance (such as $V(d|c,q) = p(rel|d,c,q)$ for IA-Select). We think that this relevance-based formulation could provide interesting properties and more flexibility (for example, to the tolerance for redundancy) to these algorithms.

5.2.4 Metrics formalization

Our unified metric framework proposed a variety of novelty and diversity metrics, some of which are based on popularity: EPC, EIP and EFD. Although each metric has a different formulation, all of them are based on the same principle of inverse popularity. The open question is to determine which of them is better, either based on a theoretic basis, an experimental approach or comparing with user feedback. Going further, a theoretical or practical discussion of the rank and relevance components (logarithmic vs. exponential rank discount, relevance with thresholds or exponentiation, etc.) could be of great use. Although we used state-of-the-art proposals, it is not fully clear that the combination of them is justified.

We plan to develop and test the generalization of further diversity metrics as described in section 4.6.3. We envision the development of user-specific discovery models, and particularizations to further contexts, such as user communities and vertical domains. This would allow us, for example, to define a notion of discovery-based distance. In addition to the provision of evaluation tools, the underlying models can be used to build objective functions for novelty and diversity enhancement methods, taking the ratings predicted by baseline recommenders as a proxy of true relevance.

Finally, we are considering the meta-evaluation of metrics as suggested by Sakai (2006) to stablish a quality criterium for metric variants regarding their novelty, rank and relevance components. As another objective criterium for the meta-evaluation of metrics, online experiments can also provide an additional way for determine the goodness of our metric framework.

Bibliography

Adomavicius, G., & Kwon, Y. (to appear). Improving Aggregate Recommendation Diversity Using Ranking-Based Techniques. *IEEE Transactions on Knowledge and Data Engineering*.

Adomavicius, G., & 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.

Agrawal, R., Gollapudi, S., Halverson, A., & Ieong, S. (2009). Diversifying search results. *2nd ACM International Conference on Web Search and Data Mining (WSDM'09)*, (pp. 5-14). Barcelona, Spain.

Anderson, C. (2006). *The Long Tail.Why the Future of Business is Selling Less of More.* Hyperion Verlag.

Bellogín, A., Cantador, I., & Castells, P. (2010). A Study of Heterogeneity in Recommendations for a Social Music Service. *1st International Workshop on Information Heterogeneity and Fusion in Recommender Systems (HetRec'10) at the 4th ACM Conference on Recommender Systems (RecSys'10)*, (pp. 1-10). Barcelona, Spain.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research, 3*(Jan), 993-1022.

Breese, J., Heckerman, D., & Kadie, C. (1998). Empirical Analysis of Predictive Algorithms for Collaborative Filtering. *14th Conference of Uncertainty in Artificial Intelligence (UAI'98)*, (pp. 43-52). Madison, WI, USA.

Carbonell, J., & Goldstein, J. (1998). The use of MMR, diversity-based reranking for reordering documents and producing summaries. *21st International ACM SIGIR Conference on Research and Developement in Information Retrieval (SIGIR'08)*, (pp. 335-336). Melbourne, Australia.

Carterette, B. (2011). An Analysis of NP-Completeness in Novelty and Diversity Ranking. *Information Retrieval, 14*(1), 89-106.

Carterette, B. (2011). System Effectiveness, User Models, and User Utility: A Conceptual Framework for Investigation. *34th International ACM SIGIR Conference on Research and Developement in Information Retrielval (SIGIR'11)*, (pp. 903-912). Beijing, China.

Carterette, B. (2011). System Effectiveness, User Models, and User Utility: A Conceptual Framework for Investigation. *34th International ACM SIGIR Conference on Research and Developement in Information Retrielval (SIGIR'11)*, (pp. 903-912). Beijing, China.

Castells, P., Vargas, S., & Wang, J. (2011). Novelty and Diversity Metrics for Recommender Systems: Choice, Discovery and Relevance. *International Workshop on Diversity in Document Retrieval (DDR'11) at the 33rd European Conference on Information Retrieval (ECIR'11).* Dublin, Ireland.

Celma, Ò., & Herrera, P. (2008). A new approach to evaluating novel recommendations. *2nd ACM Conference on Recommender Systems (RecSys'08)*, (pp. 179-186). Lausanne, Switzerland.

Chapelle, O., Metzler, D., Zhang, Y., & Grinspan, P. (2009). Expected Reciprocal Rank for Graded Relevance. *18th ACM Conference on Information and Knowledge Management (CIKM'09)*, (pp. 621-630). Singapore.

Chen, H., & Karger, D. R. (2006). Less is more: probabilistic models for retrieving fewer relevant documents. *29th International ACM SIGIR Conference on Research and Developement in Information Retrieval (SIGIR'06)*, (pp. 429-436). Seattle, WA, USA.

Clarke, C., Craswell, N., & Soboroff, I. (2009). Overview of the TREC 2009 Web Track. *18th Text Retrieval Conference (TREC 2009)*. Gaithersburg, MD, USA.

Clarke, C., Craswell, N., Soboroff, I., & Ashkan, A. (2011). A Comparative Analysis of Cascade Measures for Novelty and Diversity. *4th ACM International Conference on Web Search and Data Mining (WSDM'11)*, (pp. 75-84). Honk-Kong, China.

Clarke, C., Kolla, M., Cormack, G., Vechtomova, O., Ashkan, A., Büttcher, S., et al. (2008). Novelty and Diversity in Information Retrieval Evaluation. *31st International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'08)*, (pp. 659-666). Singapore.

Craswell, N., Zoeter, O., Taylor, M., & Ramsey, B. (2008). An Experimental Comparison of Click Position-Bias Models. *1st International Conference on Web Search and Web Data Mining (WSDM'08)*, (pp. 87-94). Palo Alto, CA, USA.

Fleder, D., & Hosanagar, K. (2009). Blockbuster Culture's Next Rise or Fall: The Impact of Recommender Systems on Sales Diversity. *Management Science*, 697-712.

Herlocker, J. L., Konstan, J. A., Terveen, L. G., & Riedl, J. T. (2004, January). Evaluating Collaborative Filtering Recommender Systems. *ACM Transactions on Information Systems, 22*(1), 5-53.

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. Y. (2011). Characterizing Search Intent Diversity into Click Models. *20th International Conference on World Wide Web (WWW'11)*, (pp. 17-26). Hyderabad, India.

Järvelin, K., & Kekäläinen, J. (2002). Cumulated Gain-Based Evaluation of IR Techniques. *ACM Transactions on Information Systems, 20*(4), 422-446.

Koren, Y., Bell, R., & Volinksy, C. (2009). Matrix Factorization Techniques for Recommender Systems. *Computer*, 30-37.

Lathia, N., Hailes, S., Capra, L., & Amatriain, X. (2010). Temporal diversity in recommender systems. *33rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'10)*, (pp. 210-217). Geneva, Switzerland.

McNee, S. M., Riedl, J., & Konstan, J. A. (2006). Being accurate is not enough: how accuracy metrics have hurt recommender systems. *Conference on Human Factors in Computing Systems (CHI'06)*, (pp. 1097-1101). Montreal, Canada.

Mei, Q., & Guo, J. (2010). DivRank: the Interplay of Prestige and Diversity in Information Networks. *16th ACM SIGKDD International Conference of Knowledge Discovery and Data Mining*, (pp. 1009-1018). Washington, DC, USA.

Moffat, A., & Zobel, J. (2008). Rank-biased precision for measurement of retrieval effectiveness. *ACM Transactions on Information Systems*, 2:1-2:27.

Radlinski, F., Kleinberg, R., & Joachims, T. (2008). Learning Diverse Rankings with Multi-Armed Bandits. *25th International Conference on Machine Learning (ICML'08)*, (pp. 784-791). Helsinki, Finland.

Robertson, S. (1977). The Probability Ranking Principle in IR. *Journal of Documentation, 33*, 294-304.

Robertson, S. (2008). A New Interpretation of Average Precision. *31st International ACM SIGIR Conference on Research and Developement in Information Retrieval*, (pp. 689-690). Singapore.

Sakai, T. (2006). Evaluating evaluation metrics based on the bootstrap. *29th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'06)*, (pp. 525-532). Seattle, WA, USA.

Santos, R., Macdonald, C., & Ounis, I. (2010). Exploiting Query Reformulations for Web Search Result Diversification. *19th International Conference on World Wide Web (WWW'10)*, (pp. 881-890). Raleigh, NC, USA.

Santos, R., Macdonald, C., & Ounis, I. (2010). Selectively Diversifying Web Search Results. *19th ACM International Conference on Information and Knowledge Management (CIKM'10)*, (pp. 1179-1188). Toronto, Canada.

Slivkins, A., Radlinski, F., & Gollapudi, S. (2010). Learning optimally diverse rankings over large document collections. *27th International Conference on Machine Learning (ICML'10)*, (pp. 983-990). Haifa, Israel.

Sweeney, S., Crestani, F., & Losada, D. (2008). "Show me more": Incremental Length Summarisation Using Novelty Detection. *Information Processing & Management*, 663-686.

Vargas, S., & Castells, P. (2011). Rank and Relevance in Novelty and Diversity Metrics for Recommender Systems. *5th ACM Conference on Recommender Systems (Recsys'11)*, (pp. 109-116). Chigago, IL, USA.

Vargas, S., Castells, P., & Vallet, D. (2011). Intent-Oriented Diversity in Recommender Systems. *34th International ACM SIGIR Conference on Research Developement in Information Retrieval (SIGIR'11)*, (pp. 1211-1212). Beijing, China.

Vargas, S., Castells, P., & Vallet, D. (2012). On the Suitability of Intent Spaces for IR Diversification. *International Workshop on Diversity in Document Retrieval (DDR'12) at the 5th ACM International Conference on Web Search and Data Mining (WSDM'12).* Seattle, WA, USA.

Voorhees, E. M., & Karman, D. K. (2005). *TREC Experiment and Evaluation in Information Retrieval.* MIT Press.

Wang, J. (2009). Mean-Variance Analysis: A New Document Ranking Theory in Information Retrieval. *31th European Conference on IR Research on Advances in Information Retrieval (ECIR'09)*, (pp. 4-16). Toulouse, France.

Wang, J., & Zhu, J. (2009). Portfolio Theory on Information Retrieval. *32nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'09)*, (pp. 115-122). Boston, MA, USA.

Yue, Y., & Joachims, T. (2008). Predicting Diverse Subsets Using Structural SVMs. *25th International Conference on Machine Learning (ICML'08)*, (pp. 1224-1231). Helsinki, Finland.

Zhai, C., Cohen, W., & Lafferty, J. (2003). Beyond independent relevance: methods and evaluation metrics for subtopic retrieval. *26th International AMC SIGIR Conference on Research and Development in Information Retrieval (SIGIR'03)*, (pp. 10-17). Toronto, Canada.

Zhang, M., & Hurley, N. (2008). Avoiding monotony: improving the diversity of recommendation lists. *2nd ACM Conference on Recommender Systems (RecSys'08)*, (pp. 123-130). Lausanne, Switzerland.

Zhang, M., & Hurley, N. (2009). Novel Item Recomendation by User Profile Partitioning. *IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technologies (WI-AT'09)*, (pp. 508-515). Milan, Italy.

Zhou, T., Kuscsik, Z., Liu, J.-G., Medo, M., Wakeling, J. R., & Zhang, Y.-C. (2010). Solving the apparent diversity-accuracy dilemma of recommender systems. *Proceedings of the National Academy of Sciences*, 4511-4515.

Ziegler, C.-N., McNee, S. M., Konstan, J. A., & Lausen, G. (2005). Improving Recommendation Lists Through Topic Diversification. *14th International Conference on World Wide Web (WWW'05)*, (pp. 22-32). Chiba, Japan.