Rank and Relevance in Novelty and Diversity Metrics for Recommender Systems

Saúl Vargas and Pablo Castells
Universidad Autónoma de Madrid
Escuela Politécnica Superior, Departamento de Ingeniería Informática
{saul.vargas,pablo.castells}@uam.es

ABSTRACT

The Recommender Systems community is paying increasing attention to novelty and diversity as key qualities beyond accuracy in real recommendation scenarios. Despite the raise of interest and work on the topic in recent years, we find that a clear common methodological and conceptual ground for the evaluation of these dimensions is still to be consolidated. Different evaluation metrics 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.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: *information filtering*.

General Terms

Algorithms, Measurement, Performance, Theory, Experimentation, Standardization.

Keywords

Novelty, diversity, metrics, evaluation, recommender systems.

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 [12,16]. 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 [7,11,20,22].

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.

The rest of the paper is organized as follows. We briefly revise the related work in the next section. The general principles of the proposed metric scheme are introduced after that in Section 3. In Section 4 we define the item novelty models upon which the metrics are built. Section 5 describes how a browsing model can be developed on the notion of choice. Model estimation methods for discovery and relevance are defined in Section 6. The relevant metric configurations resulting from these developments are presented in Section 7, with an illustrative example in Section 8.

RecSys'11, October 23–27, 2011, Chicago, Illinois, USA.

Experimental observations on the proposed metrics are reported in Section 9. We end with some final conclusions in Section 10.

2. NOVELTY AND DIVERSITY IN RECOMMENDER SYSTEMS

Novelty is a highly desirable feature for 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 [11].

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", by a specific user, or by a community as a whole. 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 global diversity over time in the user experience; and also enhances the global "diversity of sales" from the system perspective.

A common specific definition of diversity in the literature is the average pairwise dissimilarity between recommended items. Using this notion, Ziegler et al [22] define a greedy re-ranking algorithm, which diversifies baseline recommendations by iteratively selecting items that maximize a trade-off between the original recommendation value and the average distance to the new list under construction. This approach is similar to the Maximal Marginal Relevance scheme proposed in Information Retrieval (IR) for search diversification and automatic summarization [5]. The approach is evaluated by using complementary accuracy metrics (recall and precision) and studying the decrease of accuracy as diversity increases, the tradeoff being controlled by a specific parameter.

Zhang and Hurley [20] bring intra-list diversity to a more formal formulation and problem statement. Diversification is explicitly addressed as the joint optimization of two objective functions reflecting preference similarity and item diversity, which is solved by linear and quadratic programming algorithms. 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.

Recommending long-tail items, which few users have accessed to, is a common way in which novelty is understood. Zhou et al [21] define novelty as the average self-information of recommended items, which amounts to the average log inverse ratio of users who like the item (also known as "inverse user frequency"). They target this metric by means of hybrid strategies combining collaborative filtering with graph spreading techniques. Celma and Herrera [7] 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.

Lathia et al [15] take yet another angle on the diversity problem. They consider the novelty that a system delivers with respect to recommendations that it produced in the past. In a way, they measure the ability of a recommender system to evolve over time and adapt to the changing conditions of real settings.

Other authors have addressed the topic from the point of view of the recommender system, or the business behind it [11]. Adomavicius and Kwon [1] address diversity as the ability of a system to recommend as many different items as possible over the whole population –a form of aggregate diversity, defined as the union of sets of recommended items to all users in the system. The authors improve recommendations on this metric while keeping accuracy loss to a minimum, by a controlled promotion of less popular items towards the top of the recommendation rankings.

Taking on from such works, 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. Moreover, we seek specific improvements on the limitations of the metrics proposed so far. The reported metrics generally ignore the ranking of recommended items –except for the obvious application of diversity metrics at different top-n cutoffs. As a consequence, the measured diversity does not notice whether the most novel items are ranked at the top or the bottom of the recommendations. Second, the metrics do not care for the relevance of items, and focus strictly on their novelty and diversity qualities. The evaluation methodologies therefore rely on a separate accuracy metric for this purpose. We argue that it may be beneficial to handle novelty and relevance together, which is not equivalent to a combination of two separate assessments, as we shall analyze. Our view draws perspectives from the recent research on search diversity in the IR field, where diversity and accuracy are seen as two sides of the same coin that build on common principles [2,9].

3. 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 1), 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:

$$p(choose) \sim p(seen)p(rel) \qquad (1)$$

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.

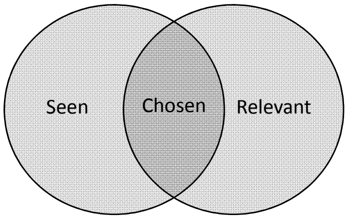

Figure 1. Discovery, choice and relevance models.

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:

$$m(R|\theta) = C\sum_{i\in R} p(choose|i,u,R)nov(i|\theta) \qquad (2)$$

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 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. 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.

4.1 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:

$$nov(i|\theta) = 1 - p(seen|i,\theta) \qquad (3)$$

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 7.3 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:

$$nov(i|\theta) = -log_2 p(seen|i,\theta) \qquad (4)$$

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:

$$nov(i|\theta) = -log_2 p(i|seen,\theta) \qquad (5)$$

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 [21]. 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.

4.2 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:

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

where $p(j|choose,\theta,i)$ is the probability that the user chooses item j in the context θ, when he 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:

$$nov(i|\theta) = \frac{\sum_{j\in\theta} p(choose|j,\theta,i)d(i,j)}{\sum_{j\in\theta} p(choose|j,\theta,i)} \qquad (6)$$

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 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 his profile–, and b) the set R of recommended items itself. In the first case, we get a user-relative novelty version of equation 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.

5. BROWSING MODEL

The browsing component of the metric scheme, as introduced in equation 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 [6,10,13,17,18], 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 he effectively gets to see *and* he finds relevant for his taste. We had already formulated this view in equation 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 he 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 [10] 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:

$$p(seen|i_k,u,R) = p(seen|i_{k-1},u,R)p(cont|k-1,u,R) = \\ = \prod_{l=1}^{k-1} p(cont|l,u,R) \qquad (7)$$

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 [17]. 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 [8], or the models in [18]. 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 [9], 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 2 can be rewritten as a configurable rank-sensitive, relevance aware metric scheme:

$$m(R|\theta) = C\sum_{i_k\in R} disc(k)p(rel|i_k,u)nov(i_k|\theta) \qquad (8)$$

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 [9], or b) the expected browsing depth, as in RBP [17] and discussed in [10]. Computing the ideal ranking is metric-specific and often costly, sometimes even NP-hard, though it can be approximated by greedy approaches [9]. 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.

6. ESTIMATION OF GROUND MODELS

6.1 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:

$$p(seen|i,r)\sim\frac{|\mathbf{i}|}{|\mathcal{U}|} = \frac{|\{u\in\mathcal{U}|r(u,i)\neq\emptyset\}|}{|\mathcal{U}|} \qquad (9)$$

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:

$$p(seen|i,\mathcal{L})\sim\frac{|\mathbf{i}|}{|\mathcal{U}|} = \frac{|\{u\in\mathcal{U}|\exists t\in\mathcal{T}:(u,i,t)\in\mathcal{L}\}|}{|\mathcal{U}|} \qquad (10)$$

$\mathcal{T}$ being the timestamp data type. Note that with these estimates, item novelty in equation 4 becomes the *inverse user frequency* IUF. The free novelty model can also be estimated over ratings or implicit data, respectively, as:

$$p(i|seen,r)\sim\frac{|\mathbf{i}|}{\sum_{j\in\mathcal{I}}|\mathbf{j}|} = \frac{|\{u\in\mathcal{U}|r(u,i)\neq\emptyset\}|}{|\{(u,j)\in\mathcal{U}\times\mathcal{I}|r(u,j)\neq\emptyset\}|} \qquad (11)$$

$$p(i|seen,\mathcal{L})\sim\frac{|\mathbf{i}|}{\sum_{j\in\mathcal{I}}|\mathbf{j}|} = \frac{|\{u\in\mathcal{U}|\exists t\in\mathcal{T}:(u,i,t)\in\mathcal{L}\}|}{|\{(u,j)\in\mathcal{U}\times\mathcal{I}|\exists t\in\mathcal{T}:(u,i,t)\in\mathcal{L}\}|} \qquad (12)$$

With the rating-based estimate (equation 11), equation 5 becomes the so-called *inverse collection frequency* ICF.

6.2 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 [8]:

$$p(rel|i,u) \sim \frac{2^{g(u,i)} - 1}{2^{g_{max}}} \qquad (13)$$

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 et al [4]. In our experiments we try a slight variation with respect to [8]: 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 [7]. Namely, $r(u,i) \longleftarrow n \cdot F(frec_{u,i})$, where $frec_{u,i}$ is the number of times u has accessed i, and $F(frec_{u,i}) \sim |\{j \in \mathbf{u} | f_{u,j} \leq f_{u,i}\}| / |\mathbf{u}|$ 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 13), this time with $\tau = 0$ – assuming that accessing an item, however infrequently, does not in general reflect a negative preference.

7. RECOMMENDATION NOVELTY AND DIVERSITY METRICS

7.1 Novelty

By plugging the popularity-based item novelty models (Section 4.1) in the general metric scheme (eq. 8), we get discovery-based recommendation novelty metrics. For instance, taking equation 3, we get:

$$nov(R|u) = \text{EPC} = C \sum_{i_k \in R} disc(k) p(rel|i_k,u)\big(1 - p(seen|i_k)\big) \quad (14)$$

which we label as expected popularity complement (EPC). Equations 4 and 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 [21].

If we take a distance-based novelty model (equation 6) relative to the set of items the target user has interacted with $\theta \equiv \mathbf{u}$ –i.e. the items in his 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):

$$nov(R|u) = \text{EPD} = C' \sum_{i_k \in R, j \in \mathbf{u}} disc(k) p(rel|i_k,u) p(rel|j,u) d(i_k,j) \quad (15)$$

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.

7.2 Diversity

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

$$div(R|u) = \text{EILD} =$$
$$= \sum_{\substack{i_k \in R \\ i_l \in R \\ l \neq k}} C_k disc(k) disc(l|k) p(rel|i_k,u) p(rel|i_l,u) d(i_k,i_l) \qquad (16)$$

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 16 thus generalizes the average intra-list distance (ILD) [20,22] with the introduction of rank-sensitivity and relevance.

Table 1. Unification of state of the art novelty and diversity metrics in the proposed metric framework.

Metric scheme	Context $\boldsymbol{\theta}$	User perspective	Generalizes
Long tail (popularity)	Ratings r or frequencies $\mathcal{L}$	Novelty	Mean self-information [21]
Distance-based	Target user u	Novelty	-
	Recommendation R	Diversity	Intra-list diversity [20,22]
Alternative discovery sources	Last recommendation $\langle u, R_{t-1} \rangle$	Novelty	Self-system diversity [15]
	All previous recommendations $\langle u, A_{t-1} \rangle$	Novelty	Self-system novelty [15]
	Recommendations by other systems $\langle u, \mathcal{S} \rangle$	Novelty	Inter-system novelty [3]
	Recommendations to other users $\langle \mathcal{U}, s \rangle$	Novelty	Inter-user diversity [3]

7.3 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 [15], 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:

$$\text{SSD}(R|u) = \frac{|R - R_{t-1}|}{|R|} \qquad (17)$$

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 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 [15], 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 [3]. Table 1 summarizes some of the metrics that can be unified in our framework by different instantiations of θ in the item novelty scheme.

8. 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 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 his 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 2. 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.

Table 2. An illustrative example.

	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

Based on equations 9 and 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 3. The best result is underlined 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$_{\text{rel}}$ 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$_{\text{rank}}$ 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$_{\text{rank,rel}}$ is the metric that best discriminates this fact.

Table 3. Resulting values of different metrics for the two example recommendations, combining different rank and relevance configurations in the EPC novelty metric.

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

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$_{\text{rank,rel}}$ does not suffer from these shortcomings.

9. 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 Ò. Celma [7], 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 do 5-fold cross-validation on five 80-20% random training-test rating splits. 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.

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 [14] 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 [7], 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 4.

Table 4. Accuracy of the tested baselines, measured in nDCG@50 over the two datasets.

	CB	MF	UB	AVG	RND
MovieLens 1M	0.1113	0.2136	0.1463	0.1497	0.0332
Last.fm	-	0.3081	0.5797	0.0392	0.0107

The discovery models (equations 3-5) are built on training data –since they do not involve target users– and the relevance models (equation 13) on test data. The estimation of the discovery models is based on equations 9 and 11 for MovieLens (explicit ratings) and equations 10 and 12 for Last.fm (item access log). The browsing models build exclusively on test data (for relevance, equation 13)

Table 5. 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 [17], 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).

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

and recommenders' output (for recommendation discovery distribution, equation 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.

9.1 Pure and Relevance-aware Metrics

Figure 2 shows how the tested recommenders compare on different metrics, namely EPC, EPD, and EILD (equations 14, 15, 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 [7]), 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 [20,21,22].

The two bottom graphics in Figure 2 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 4), 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.

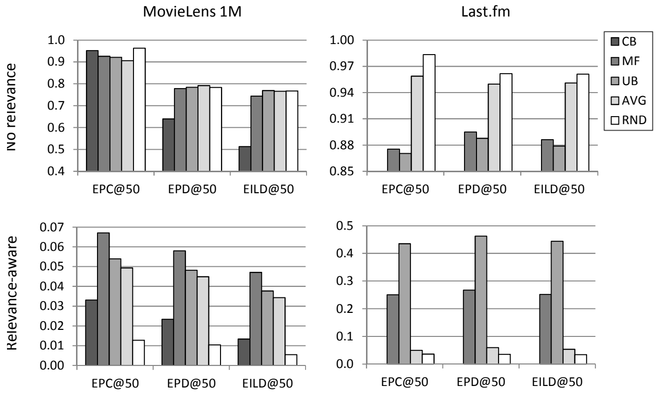

Figure 2. Novelty and diversity metrics are shown on four baselines (content-based, matrix factorization, user-based kNN, average, and random) over MovieLens 1M –two graphics on the left– and Last.fm –right. The top two graphics display metrics that ignore relevance, whereas the bottom ones are relevance-aware. All the metrics in the figure are rank-insensitive.

9.2 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 [22], which we term Maximal Marginal Relevance (MMR) for its connection to the approach described in [5], 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) is used in place of the dissimilarity component; and c) an adaptation of the IA-Select algorithm [2], originally devised for search diversification (see [19] for more details on this adaptation). Additionally, we include a random re-ranking.

Table 5 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 [17]) 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.

10. 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.

The directions to continue the research presented here are manifold. We plan to develop and test the generalization of further diversity metrics as described in Section 7.3. We envision the development of user-specific discovery models, and particularizations to further contexts, such as user communities and vertical domains. 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, but not least, we see the connection to the recent work on search diversity in the IR field as a relevant future research direction. The diversity problem is being stated in the IR community as an issue of query ambiguity and underspecification, which is formulated in terms of query aspects, interpretations, or similar notions [2,9]. Expressing our relevance model in terms of some analogous notion of item aspects is straightforward; the main difficulty lies in the right conception and identification of such aspects in items and user profiles –first steps in this direction are reported in [19]. The introduction of aspects in the discovery models is less direct in comparison, and worthy of exploration as well.

11. ACKNOWLEDGMENTS

This work is supported by the Spanish Government (TIN2011-28538-C02-01), and the Government of Madrid (S2009TIC-1542).

12. REFERENCES

[1] Adomavicius, G. and Kwon, Y. Improving Aggregate Recommendation Diversity Using Ranking-Based Techniques. *IEEE Transactions on Knowledge and Data Engineering*. In Press.

[2] Agrawal, R., Gollapudi, S., Halverson, A., and Ieong, S. Diversifying search results. WSDM 2009, Barcelona, Spain, 5-14.

[3] Bellogín, A., Cantador, I., and Castells, P. A Study of Heterogeneity in Recommendations for a Social Music Service. HetRec Workshop at RecSys 2010, Barcelona, Spain, 1-10.

[4] Breese, J. S., Heckerman, D., and Kadie, C. Empirical analysis of predictive algorithms for collaborative filtering. UAI 1998, Madison, WI, USA, 43-52.

[5] Carbonell, J. G. and Goldstein, J. The Use of MMR, Diversity-Based Reranking for Reordering Documents and Producing Summaries. SIGIR 1998, Melbourne, Australia, 335-336.

[6] Carterette, B. System effectiveness, user models, and user utility: a conceptual framework for investigation. SIGIR 2011, Beijing, China, 903-912.

[7] Celma, O. and Herrera, P. A New Approach to Evaluating Novel Recommendations. RecSys 2008, Lausanne, Switzerland, 179-186.

[8] Chapelle, O., Metzler, D., Zhang, Y., and Grinspan, P. Expected Reciprocal Rank for Graded Relevance. CIKM 2009, Hong Kong, China, 621-630.

[9] Clarke, C. L. A., Kolla, M., Cormack, G. V., Vechtomova, O., Ashkan, A., Büttcher, S., and MacKinnon, I. Novelty and diversity in information retrieval evaluation. SIGIR 2008, Singapore, 659-666.

[10] Clarke, C. L. A., Craswell, N., Soboroff, I., and Ashkan, A. A Comparative Analysis of Cascade Measures for Novelty and Diversity. WSDM 2011, Hong-Kong, China, 75-84.

[11] Fleder, D. M. and Hosanagar, K. Blockbuster Culture's Next Rise or Fall: The Impact of Recommender Systems on Sales Diversity. *Management Science* 35, 5, 2009, 697-712.

[12] Herlocker, J. L., Konstan, J. A., Terveen, L. G., and Riedl, J. T. Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems* 22, 1, 2004, 5-53.

[13] Hu, B., Zhang, Y., Chen, W., Wang, G., and Yang, Q. Characterizing Search Intent Diversity into Click Models. WWW 2011, Hyderabad, India, 17-26.

[14] Koren, Y., Bell, R., Volinsky, C. Matrix Factorization Techniques for Recommender Systems. *IEEE Computer* 42, 8, 2009, 30-37.

[15] Lathia, N., Hailes, S., Capra, L., Amatriain, X. Temporal Diversity in Recommender Systems. SIGIR 2010, Geneva, 210-217.

[16] McNee, S. M., Riedl, J., and Konstan, J. A. Being Accurate is Not Enough: How Accuracy Metrics have hurt Recommender Systems. CHI 2006, Montréal, Canada, 1097-1101.

[17] Moffat, A. and Zobel, J. Rank-biased precision for measurement of retrieval effectiveness. *ACM Transactions on Information Systems* 27, 1, 2008.

[18] Radlinski, F., Kleinberg, R., Joachims, T. Learning diverse rankings with multi-armed bandits. ICML 2008, Helsinki, Finland, 784-791.

[19] Vargas, S., Castells, P., and Vallet, D. Intent-Oriented Diversity in Recommender Systems. SIGIR 2011, Beijing, China, 1211-1212.

[20] Zhang, M. and Hurley, N. Avoiding Monotony: Improving the Diversity of Recommendation Lists. RecSys 2008, Lausanne, Switzerland, 123-130.

[21] Zhou, T., Kuscsik, Z., Liu, J-G., Medo, M., Wakeling, J. R., and Zhang, Y-C. Solving the apparent diversity-accuracy dilemma of recommender systems. *Proc. of the National Academy of Sciences of the United States of America* 107, 10, 2010, 4511-4515.

[22] Ziegler, C-N., McNee, S. M., Konstan, J. A., and Lausen, G. Improving recommendation lists through topic diversification. WWW 2005, Chiba, Japan, 22-32.