Novelty and Diversity Metrics for Recommender Systems: Choice, Discovery and Relevance

Pablo Castells[1], Saúl Vargas[1], and Jun Wang[2]

[1] Departamento de Informática, Universidad Autónoma de Madrid, 28049 Spain
[2] Department of Computer Science, University College London, WC1E 6BT, UK

{pablo.castells,saul.vargas}@uam.es, wang.jun@acm.org

Abstract. There is an increasing realization in the Recommender Systems (RS) field that novelty and diversity are fundamental qualities of recommendation effectiveness and added-value. We identify however a gap in the formalization of novelty and diversity metrics –and a consensus around them– comparable to the recent proposals in IR diversity. We study a formal characterization of different angles that RS novelty and diversity may take from the end-user viewpoint, aiming to contribute to a formal definition and understanding of different views and meanings of these magnitudes under common groundings. Building upon this, we derive metric schemes that take item position and relevance into account, two aspects not generally addressed in the novelty and diversity metrics reported in the RS literature.

Keywords: novelty, diversity, metrics, evaluation, recommender systems

1 Introduction

Several approaches to assess novelty and diversity in search results have been proposed in the last few years [1,5,7,8,10]. Datasets have been released at evaluation campaigns such as TREC 2009/10, fostering convergence and sharing of common benchmarks. Metrics such as α-nDCG, nDCG-IA, MAP-IA, ERR-IA, NRBP have been used in the diversity task of the TREC Web track [6]. Further diversity-related metrics have been proposed outside this, such as subtopic precision and recall [5] or k-call [10]. Even though diversity and novelty are a largely open research topic in IR today, one can see a fair extent of convergence and reuse of metrics and methodologies in the community.

In contrast, studies of comparable depth on measuring novelty and diversity, and/or an array of well-understood metrics, are still missing in the Recommender Systems (RS) area. In fact, the range of metrics described in the literature is considerably scant. For instance, to the best of our knowledge, a measure that takes into account the order of recommended items is completely missing –except for the obvious application of diversity metrics at different top-n cutoffs. Yet novelty and diversity play an arguably even more central role in the recommendation context, where the practical value and gain from recommendation are closely linked to the notion of discovery in most scenarios. Moreover, the ambiguity in user needs is considerably higher than in ad-hoc IR, and intrinsic to the task, since there is no explicit expression of such needs. Despite a significant and growing stream of research and interest in diversity and novelty in the field [11,12,13], there would seem to be a gap in the definition and systematic study of metrics.

Following this motivation, we discuss the definition of suitable metrics for the needs and specifics of RS regarding novelty and diversity. Our study considers two main ground concepts in recommendation novelty, namely item *similarity* and user-item *interaction*, upon which different recommendation novelty and diversity models unfold. User-item interaction is in turn modeled upon three core conditions: *choice*, *dis-*

covery, and *relevance*. As a specific result, we find modular means to introduce rank and relevance sensitiveness in the metrics, two properties currently not present in the diversity and novelty metrics reported in the RS literature.

2 Novelty and Diversity in Recommender Systems

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. Another fundamental take on diversity is defined in ad-hoc IR in terms of query interpretations or aspects. The adaptation of this perspective to a recommendation task certainly deserves investigation. For reason of available space, we leave it aside in the present paper, and we only discuss here –besides novelty itself– notions of diversity that result from a novelty model, as we shall see. Moreover, we focus on novelty and diversity as perceived by the end-user, i.e. we do not cover here the system or the business perspective. Finally, we assume an application scenario where the items that the user has already chosen in the past are not recommended again –leaving out scenarios such as recommendation for grocery shopping (where the same products are bought periodically), or personalized music playlist generation (where it is generally ok to recommend known music tracks).

We distinguish two main notions upon which recommendation novelty and diversity can be defined: item popularity and similarity. Recommendation novelty and diversity can be modeled upon the novelty and dissimilarity of recommended items, which in turn we formalize in terms of user-item interaction models, and distance functions. Novelty and diversity can be measured generically, that is, irrespective of the user they are delivered to, or they can actually take into account the target user. In a generic approach, the diversity in a list of items can be measured, for instance, in terms of the objective variety of items in the list (e.g. as pairwise dissimilarity), and novelty can be defined in terms of how many users are familiar with the items. In a user-relative approach, novelty can take into account what the specific target user has already seen, and diversity can consider the variety of interests within his individual user profile. Metrics may just analyze the composition of recommended lists, or they may also take into account that the top positions have a higher impact on the effective diversity and novelty value of the list. A metric may strictly focus on novelty, leaving relevance for a complementary metric to capture it, or actually require items to be relevant for their novelty to be counted in. Which among all such variants is more appropriate depends on the evaluation goals and requirements, the specifics of the recommendation task and/or the application domain.

3 Item Novelty Models

We consider two models of item novelty, one oriented to popularity, and one based on inter-item distance. We consider two variants for each formulation: generic and relative to an item set. In general we will consider two items sets in relative novelty: the profile of the target user, and (in distance-based models) a list of recommended items. Each will induce different recommendation metrics, as we shall see later on in Section 5.

Popularity-based Item Novelty. The novelty of an item can be defined relative to a set of observed events on the set of all items. A common way to formalize the *generic novelty* of an item is by the amount of information its observation conveys [12], in terms of some distribution involving the item. This is expressed in Information Theory as:

$$novelty(i) = I(i) = -\log_2 p(i) \tag{1}$$

where $p(i)$ represents the probability that i is observed, and $I(i)$ is commonly called self-information or surprisal. In this model, we propose to interpret the i random variable as an event of user choice, that is, "i is picked" by a random user. This reflects a factor of item popularity, whereby $novelty(i)$ corresponds to the log of the *inverse popularity*. High novelty values correspond to long-tail items in the density function, that few users have chosen or interacted with, and low novelty values correspond to popular head items. This scheme measures generic novelty as far as it is the same for all users. A *user-relative novelty* variant can be defined by simply taking $p(i|u)$ in equation 1, which amounts to restricting our observations to the target user:

$$novelty(i|u) = -\log_2 p(i|u) \tag{2}$$

An alternative, discovery-based popularity model is to consider the probability $p(K|i)$ that an item is known or is familiar to (rather than chosen by) a random user. In this case, we define generic and user-relative novelty respectively as:

$$novelty(i) = 1 - p(K|i) \qquad novelty(i|u) = 1 - p(K|i,u) \tag{3}$$

In order to emphasize the effect of highly novel items (favoring few very novel items vs. many moderately novel), one may also consider the logarithm of the inverse probability:

$$novelty(i) = -\log_2 p(K|i) \qquad novelty(i|u) = -\log_2 p(K|i,u) \tag{4}$$

Distance-based Item Novelty. Relative novelty can also be modeled with respect to a set of items on a Euclidean view. This can be defined as the *average* or *minimum distance* between the item at hand, and the items in the set:

$$novelty(i|S) = \sum_{j \in S} p(j|S)d(i,j) \quad \text{or} \quad novelty(i|S) = \min_{j \in S} d(i,j) \tag{5}$$

where d is some distance measure. The distance can be defined e.g. as $d(i,j) = 1 - sim(i,j)$ for 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). If we take S as the set of items a user has interacted with (i.e. the items in his profile), we get a user-relative novelty version of equation 5.

In the next section we discuss several estimation approaches for the distributions that have come up so far. $p(i|S)$ will be discussed only in the case where S is a user profile. In section 5 we discuss the case where $S = R$, a list of recommended items.

4 Ground Models

The models upon which novelty is defined in the previous section can use different estimation approaches, depending on the availability and type of observation data, the choice of random variables and any additional restriction on the observed events upon which the distributions are estimated. We broadly distinguish three main categories of user-item relationships:

- *Choice*: an item is used, picked, selected, accessed, browsed, bought, etc. It is common to have a frequency associated to this event, though the relation can also be binary (e.g. one-time purchase).

- *Discovery*: an item has/has not been seen before. This is understood as a binary fact, independently from the frequency of interaction, or the degree of enjoyment / dislike.
- *Relevance*: in the context of RS, relevance can be related to notions of preference, i.e. how much a user likes or enjoys an item, or how useful the item is.

Choice and discovery aspects in the interaction between users and items gives rise to different novelty and diversity model variants, which can be implemented in different ways depending on the available data, as we discuss next. We do not see relevance as playing a role in the popularity models discussed in the previous section, but we also discuss it here, as another ground aspect of user-item interaction modeling, which we shall use later on in recommendation metrics. Choice models are most naturally associated to observations in the form of usage data, whereas relevance models are best estimated upon explicit user ratings, and both types of observation suit discovery modeling well. We nonetheless discuss estimation approaches for choice in terms of ratings as well, and relevance in terms of usage. As a general rule, choice and discovery model estimates should use training data only –preferably separate from the training data used by the recommender–, whereas relevance estimates should use test data.

Item Choice. As a simple abstraction for observed usage, let us assume the observed data consists of a set $\mathcal{L}$ of user/item/timestamp records, reflecting item access by users (e.g. in an online music site). Taking $p(i)$ as the probability that i is used, a maximum likelihood item prior estimate would be:

$$p(i) \sim \frac{|\{(u,i,t) \in \mathcal{L}\}|}{|\mathcal{L}|} \tag{6}$$

Under this formulation, novelty as defined in equation 1 is the so-called *inverse collection frequency* ICF of the item. The posterior $p(i|u)$ for user-relative novelty (equation 2) is trickier as far as we should assume no observation of u accessing i in the past (as stated in the introduction, we otherwise assume i would not be recommended to u). We can take an indirect estimate based on other items the user has accessed:

$$p(i|u) \sim \sum_{j \in \mathbf{u}} p(i|j)p(j|u), \quad p(j|u) \sim \frac{|\{(u,j,t) \in \mathcal{L}\}|}{|\{(u,k,t) \in \mathcal{L}\}|} \text{ if } j \in \mathbf{u}, \quad p(i|j) \sim \frac{|\mathbf{i} \cap \mathbf{j}|}{|\mathbf{j}|} \tag{7}$$

where $\mathbf{u}$ denotes the set of items in the user's profile, and $\mathbf{i}$ the set of users who have accessed i. User ratings are sparse observations to support choice models, but can still enable an acceptable rough estimation if enough data are available. This can be done by equating a positive rating (i.e. above some threshold τ) to a one time access observation (as in e.g. a purchase), which fits as a model for equations 6 and 7 above (see Table 1).

Item Discovery. The prior that a random user knows about i can be estimated as:

$$p(K|i) \sim \frac{|\mathbf{i}|}{|\mathcal{U}|} = \frac{|\{u \in \mathcal{U} | \exists t \in \mathcal{T}: (u,i,t) \in \mathcal{L}\}|}{|\mathcal{U}|} \tag{8}$$

where $\mathcal{U}$ is the set of all users, and $\mathcal{T}$ is the timestamp data type. In this formulation, item novelty in equation 4 becomes the *inverse user frequency* IUF [2]. When the observed data consists of item ratings by users, this becomes:

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

where $r(u,i) \neq \emptyset$ means the rating of u for i is known. If the interaction between users and items is binary (e.g. one-time purchase), then equations 8 and 9 are the same. To model $p(K|i,u)$ in user-relative novelty, assuming again no past interaction between u and i, we can take an indirect estimate: $p(K|i,u) \sim \sum_{j \in \mathbf{u}} p(K|j)p(i|j)\, p(j|u)/p(i|u)$.

Item Relevance. Relevance in RS can be equated to the user interest for items. How relevance can be modeled depends again on the nature of available observations. For usage logs, a correspondence can be fairly established between item usage counts and user interest, in such a way that probability estimates of an item being used –$p(i|u)$– can be (properly scaled and) taken as a reasonable proxy for the probability of the item being liked (i.e. relevant). Under this view, the approaches discussed above for item choice (eq. 6 and 7) would apply here. If instead the available input consists of explicit user ratings, the probability of items being liked can be modeled by some heuristic mapping between rating values and probability of relevance. For instance, drawing from the ERR metric scheme [4]:

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

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 proposed by Breese et al [2].

The estimation approaches described here thus provide complete means to instantiate and compute the novelty models defined in the preceding subsection, in terms of item/user frequencies and/or rating data. Table 1 below summarizes some of the combinations that result from the alternatives discussed so far.

Table 1. Summary of item novelty models (smoothing to be applied in the estimates as appropriate), where $\mathcal{I}$ denotes the set of all items.

			Model estimation	
	Novelty model	Approach	Usage data	Rating data
Generic	Item choice (ICF)	$-\log_2 p(i)$	$p(i) \sim \frac{\lvert\{(u,i,t) \in \mathcal{L}\}\rvert}{\lvert\mathcal{L}\rvert}$	$p(i) \sim \frac{\lvert\{u \in \mathcal{U} \mid r(u,i) > \tau\}\rvert}{\lvert\{(u,j) \in \mathcal{U} \times \mathcal{I} \mid r(u,j) > \tau\}\rvert}$
Generic	Item discovery (IUF)	$1 - p(K\|i)$ $-\log_2 p(K\|i)$	$p(K\|i) \sim \frac{\lvert\{u \in \mathcal{U} \mid \exists t \in \mathcal{T}: (u,i,t) \in \mathcal{L}\}\rvert}{\lvert\mathcal{U}\rvert}$	$p(K\|i) \sim \frac{\lvert\{u \in \mathcal{U} \mid r(u,i) \neq \emptyset\}\rvert}{\lvert\mathcal{U}\rvert}$
Relative	Item choice	$-\log_2 p(i\|u)$	$p(i\|u) \sim \sum_{j \in \mathbf{u}} \frac{\lvert\mathbf{i} \cap \mathbf{j}\rvert}{\lvert\mathbf{j}\rvert} \frac{\lvert\{(u,j,t) \in \mathcal{L}\}\rvert}{\lvert\{(u,k,t) \in \mathcal{L}\}\rvert}$	$p(i\|u) \sim \sum_{j \in \mathbf{u}} \frac{\lvert\mathbf{i}_\tau \cap \mathbf{j}_\tau\rvert}{\lvert\mathbf{j}_\tau\rvert \lvert\mathbf{u}_\tau\rvert}$ | $\mathbf{u}_\tau = \{i \in \mathcal{I} \mid r(u,i) > \tau\}$, $\mathbf{i}_\tau = \{u \in \mathcal{U} \mid r(u,i) > \tau\}$
Relative	Item discovery	$1 - p(K\|i,u)$ $-\log_2 p(K\|i,u)$	$p(K\|i,u) \sim \sum_{j \in \mathbf{u}} p(K\|j) \frac{\lvert\mathbf{i} \cap \mathbf{j}\rvert}{\lvert\mathbf{j}\rvert} p(j\|u)/p(i\|u)$	
Relative	Avg. user distance	$\frac{1}{\lvert u\rvert} \sum_{j \in \mathbf{u}} d(i,j)$	–	–
Relative	Min. user distance	$\min_{j \in \mathbf{u}} d(i,j)$	–	–

5 Recommendation Novelty and Diversity Metrics

As a general scheme, we define metrics on recommender systems' output as the expected novelty of the recommended items:

$$m(R) = \sum_{i \in R} p(i|R)\, novelty(i) \tag{10}$$

where R is the list of recommended items. An interesting user-relative derivation of this formulation consist in modeling $p(i|R,u)$ by considering that a user u picks item i if a) the user browses as far as the position of i in the ranking, and b) he decides to pick i because he is interested in it (relevance). If we assume that both facts are independent, based on a generic user model, who at each position k in the ranking continues browsing down to the next position with some probability $p(k)$ (as modeled in [9]), we get:

$$m(R|u) = \sum_n \left(\prod_{k<n} p(k) \right) p(rel|i_n, u)\, novelty(i_n|u)$$

where i_n is the item at position n in R and $p(rel|i_n,u)$ is the probability that u finds i_n

relevant. The term $p(rel|i_n, u)$ thus introduces a condition of relevance: the potential novelty of i_n shall be counted in the overall novelty assessment as much as the user possibly likes the item. This is in contrast with the metrics in the RS literature, which focus on novelty or diversity only, and require a complementary accuracy metric to capture relevance. Furthermore, $p(k)$ introduces a component that makes the metric rank-sensitive. For instance, similar to the RBP scheme [9], we may consider a constant $p(k) = p$ for all k, thus getting an exponential rank discount: $(R|u) = (1-p)\sum_n p^{n-1} p(rel|i_n, u)$ $novelty(i_n|u)$. Other heuristic discount schemes can be considered as well, such as a logarithmic discount (as in nDCG), a linear discount, etc. The general form would thus be:

$$m(R|u) = \sum_n disc(n) p(rel|i_n, u) novelty(i_n|u) \tag{11}$$

where $disc(n)$ is the discount function. The discount, the relevance term, and the user-dependencies can be included or excluded as best fits the evaluation requirements.

The schemes discussed so far apply to all the item novelty models defined in Section 3. In general, popularity and user-relative distance give rise to recommendation novelty metrics, whereas distance-based set-relative novelty with respect to R results in recommendation diversity metrics, as we discuss next.

Recommendation Novelty Metrics. Introducing the simple self-information (choice-oriented) item novelty model defined in equation 1 into equation 10, we get:

$$novelty(R) = -\sum_{i \in R} p(i|R) \log_2 p(i)$$

which gives a measure of overall recommendation novelty. Under an item choice model, this can be read as the expected ICF of the recommended items. Further, if we make the approximation $p(i|R) \sim p(i)\,|\mathcal{I}|/|R|$ for $i \in R$ ($\mathcal{I}$ being the set of all items), it can be seen that this turns out to be $novelty(R) \sim H(R) + C$, which is (except for a constant $C = \log_2 \frac{|\mathcal{I}|}{|R|}$) the entropy of R under the $p(\cdot\,|R)$ distribution, a common RS novelty metric [12]. Alternatively, the rank-sensitive, relevance-aware development, and the user-specific variants discussed above (equation 11) may apply, whereby we get enhanced alternatives to the plain entropy metric.

Using the discovery-oriented popularity model defined by equation 4, we get:

$$novelty(R) = -\sum_{i \in R} p(i|R) \log_2 p(K|i)$$

which corresponds to the expected IUF of the recommended items. Using equation 3 instead of 4, we get $novelty(R) = \sum_{i \in R} p(i|R)\big(1 - p(K|i)\big)$, the expected probability that an item in the recommended list is not known by the user. This can also be read as the expected number of unknown items in the recommendation, a natural and direct measure of novelty. Again, relevance and rank bias can be introduced by refining $p(i|R,u)$ into $disc(n)p(rel|i_n, u)$ in the user-relative discovery-oriented formulations.

If we take a distance-based user-relative novelty model (equation 5 with $S = \mathbf{u}$), starting from equation 11, we get an alternative novelty measure consisting of the expected distance between the recommended items and the items in the user profile:

$$novelty(R|u) = \sum_{n, j \in \mathbf{u}} disc(n) p(rel|i_n, u) p(j|u) d(i, j)$$

where $p(j|u)$ can be e.g. simplified to a uniform distribution, or be understood as an additional relevance factor, equated to $p(rel|j, u)$, in which case item relevance would be twice accounted for in the metric.

Novelty-based Diversity Metrics. Taking on from equation 11, and instantiating set-relative distance-based novelty models (eq. 5) with $S = R$, we get a measure of diversity:

$$
\begin{aligned}
diversity(R|u) &= \sum_{n,k} disc(n)p(rel|i_n, u)p(i_k|R)d(i_n, i_k) \\
&= 2\sum_{k<n} disc(n)disc(k)p(rel|i_n, u)d(i_n, i_k)
\end{aligned}
\qquad (12)
$$

This general form provides a rank-sensitive and doubly rank-aware *expected intra-list diversity* metric (where assuming d is symmetric and since $d(i,i) = 0$, it is enough to sum for $k < n$). Equation 12 generalizes the average intra-list distance –used in several works on recommendation diversity [11,13]– with the introduction of rank-sensitivity and relevance. Again, the discount and relevance factors can be included or excluded as best fits the evaluation requirements. In particular, if we simplify the discount factors to uniform priors at each raking position (no discount is applied), and relevance is not considered in the model, equation 12 reduces to plain average intra-list diversity: $diversity(R|u) = \frac{2}{|R|(|R|-1)}\sum_{k<n} d(i_n, i_k)$, as used in the RS literature.

6 Experimental Results

Table 2 shows the value of particular instantiations of the above metric schemes in an experiment with MovieLens 100K data. The metrics apply to a common state of the art kNN collaborative filtering recommender (user-based with 50 neighbors), and three diversification algorithms on the baseline output, which rerank the top 500 items based on three diversification algorithms: an adaptation of IA-Select [1], two MMR schemes [3] (with diversity components based on movie genre similarity and IUF, respectively, both tuned towards high diversity with $\lambda = 0.6$), and a random reranking.

Table 2. Sample results for three representative metric schemes (generic novelty, user-relative novelty, generic diversity), in four configurations: rank and relevance insensitive (None), rank-sensitive (Rank), relevance-aware (Rel), and Both. Values better than random are in bold, italics indicate above the kNN baseline, and the best value for each metric is underlined. All differences are statistically significant (Wilcoxon p < 0.01) except when in parenthesis (w.r.t. random) and brackets (kNN).

	Expected IUF (EIUF@50)				Expected profile dist. (EPD@50)				Expected ILD (EILD@50)			
	None	Rank	Rel	Both	None	Rank	Rel	Both	None	Rank	Rel	Both
kNN	3.3815	(3.4149)	**0.2108**	**0.2178**	0.8289	0.8303	**0.0529**	**0.0541**	0.7944	0.4812	**0.0507**	**0.0323**
IA-Select	3.1983	3.2753	**0.1814**	**0.1918**	***0.8707***	***0.8630***	**0.0510**	**0.0519**	***0.8836***	***0.5379***	***0.0516***	***0.0331***
MMR-dist	3.3598	(***3.4487***)	**0.2065**	**0.2162**	***0.8666***	***0.8733***	***0.0545***	***0.0559***	***0.8900***	***0.5453***	***0.0562***	***0.0360***
MMR-iuf	***4.5297***	***4.8009***	***0.2478***	***0.2649***	*0.8319*	(*0.8344*)	**0.0467**	**0.0472**	[0.7917]	[0.4795]	**0.0450**	**0.0282**
Random	*3.4326*	[*3.4396*]	0.1726	0.1729	*0.8371*	*0.8370*	0.0436	0.0436	*0.8268*	*0.5004*	0.0439	0.0275

It can be seen how the different metrics capture different aspects of recommendations. Random reranking beats the baseline in all of the relevance-unaware metrics (and is even second best on EIUF), an effect that is consistently reversed with the introduction of relevance. Relevance also reveals an above-random performance by IA-Select and MMR-dist on EIUF (unnoticed by the relevance-unaware variant). Note to this respect that EIUF and EILD in the "None" variant correspond with approaches reported in the RS literature. MMR-dist is best at relevance except for EIUF, where MMR-iuf is best, as one would expect, as it greedily targets IUF. It can also be seen that rank sensitivity uncovers a better performance by MMR-dist over the baseline and random reranking on EIUF without relevance –which is not perceived when disregarding the ranking. It also shows that while IA-

Select is slightly better than MMR-dist at pure EPD regardless of item order, MMR-dist ranks the novel items better. The overall effect of rank is less significant than relevance in this experiment though. This is because the diversifiers rerank the top 500 items, while the metrics take a fairly shorter top 50 cutoff, thereby capturing rank improvement to some extent even in the rank-unaware variants. Experiments with different baselines and configurations (which we omit here for lack of space) confirm and extend our observations.

7 Conclusion

The presented study aims to contribute to the understanding of the different perspectives on novelty –and derived diversity– in RS, laying out the different views, alternatives, variants, and means of estimation, upon a common, formalized ground. Our effort aims to cover and generalize the metrics reported in the RS literature [11,12,13], and derive new ones. Two novel features in novelty and diversity measurement arise from our study: ranking sensitivity, and relevance-awareness. Both aspects are introduced in a generalized way by easy to configure terms in any metric supported by our scheme. Preliminary experiments confirm our hypotheses and provide initial observations on the behavior of the different metric configurations. Room remains for deeper examination, and additional empirical studies in specific tasks and scenarios to provide further insights on the qualities of the metrics for different purposes.

Acknowledgments. This work was supported by the Spanish Ministry of Science and Innovation (TIN2008-06566-C04-02) and the Government of Madrid (S2009/TIC-1542).

References

1. R. Agrawal, S. Gollapudi, A. Halverson, and S. Ieong. Diversifying search results. WSDM 2009, Barcelona, Spain, pp. 5-14.
2. J. S. Breese, D. Heckerman, and C. Kadie. Empirical analysis of predictive algorithms for collaborative filtering. UAI 1998, Madison, WI, USA, pp. 43-52.
3. J. G. Carbonell, J. Goldstein. The Use of MMR, Diversity-Based Reranking for Reordering Documents and Producing Summaries. SIGIR 1998, Melbourne, Australia, pp. 335-33.
4. O. Chapelle, D. Metzler, Y. Zhang, and P. Grinspan. Expected Reciprocal Rank for Graded Relevance. CIKM 2009, Hong Kong, China, pp. 621-630.
5. H. Chen and D. R. Karger. Less is More. SIGIR 2006, Seattle, WA, USA, pp. 429-436.
6. C. L. A. Clarke, N. Craswell, and I. Soboroff. Overview of the TREC 2009 Web Track. TREC 2009, Gaithersburg, MD, USA.
7. C. L. A. Clarke, M. Kolla, G. V. Cormack, O. Vechtomova, A. Ashkann, S. Büttcher, and I. MacKinnon. Novelty and diversity in information retrieval evaluation. SIGIR 2008, Singapore, pp. 659-666.
8. C. L. A. Clarke, M. Kolla, and O. Vechtomova. An Effectiveness Measure for Ambiguous and Underspecified Queries. ICTIR 2009, Cambridge, UK, pp. 188-199.
9. A. Moffat and J. Zobel. Rank-biased precision for measurement of retrieval effectiveness. ACM Transactions on Information Systems 27(1), December 2008.
10. C. Zhai, W. W. Cohen, and J. Lafferty. Beyond independent relevance: Methods and evaluation metrics for subtopic retrieval. SIGIR 2003, Toronto, Canada, pp. 10-17.
11. M. Zhang and N. Hurley. Avoiding Monotony: Improving the Diversity of Recommendation Lists. RecSys 2008, Lausanne, Switzerland, pp. 123-130.
12. T. Zhou, Z. Kuscsik, J-G. Liu, M. Medo, J. R. Wakeling, and Y-C. Zhang. Solving the apparent diversity-accuracy dilemma of recommender systems. Proceedings of the National Academy of Sciences of the United States of America 107(10), 2010, pp. 4511-4515.
13. C-N. Ziegler, S. M. McNee, J. A. Konstan, G. Lausen. Improving recommendation lists through topic diversification. WWW 2005, Chiba, Japan, pp. 22-32.