New Approaches to Diversity and Novelty in Recommender Systems

Saúl Vargas
Universidad Autónoma de Madrid
Departamento de Ingeniería Informática
Madrid, 28049 Spain
saul.vargas@uam.es

Beyond accuracy, novelty and diversity have attracted increasing interest as quality factors of Recommender Systems (RS) in the last few years. This paper presents work in progress towards the application of intent-oriented IR diversity techniques to the RS field, and the formalization of novelty and diversity metrics for RS. Experimental results show that the resulting diversification techniques produce interesting results with respect to the metrics defined here.

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

1. INTRODUCTION

Recommender Systems (RS) can be seen as a particular case of personalized Information Retrieval in which 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 system 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 user 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 himself –even if the user happens to like the items, the chances that he 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 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 effectiveness.

The problem of results diversity has been already addressed in standard IR, but to solve query disambiguation or underspecification. Current algorithms 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, and promoting less widely known (so-called long-tail) items in recommendations. 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. 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.

We present here our work in progress in these directions. On the one hand, we explore the adaptation of IR diversity perspectives to RS, seeking equivalences between the principles proposed in search diversity to the elements involved in a recommendation scenario. On the other hand, we research the development of a framework for the definition novelty an diversity metrics that unifies different perspectives, 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.

2. RELATED WORK

Search result diversity is being actively researched in the IR field as a means to address query ambiguity and underspecification, creating new retrieval algorithms, re-ranking the best results of a previous IR systems or just creating new metrics. Carbonell et al. (1998) defined a re-ranking algorithm called Maximal Marginal Relevance for maximizing the difference to the previous ranked search results whilst maintaining the relevance. Clarke et al. (2008) presented a version of NDCG to evaluate both diversity and accuracy. Agrawal et al. (2009) used a taxonomy of intent-aware information to diversify search results through re-ranking and under the same framework proposed intent-aware variants of some state-of-art metrics. Santos et al. (2010) introduced a novel probabilistic framework for Web search result diversification. Capannini et al. (2010) used query logs for extract query interpretations and diversifying the original results. Mei et al. (2010) proposed DivRank, a random walk algorithm for balancing prestige and diversity on a network of linked documents.

Agrawal et al. (2009) and Santos et al. (2010) materialize the notion of query intent as explicit "subtopics" –or broadly equivalent terminology such as sub-queries, query aspects, etc.–, which are associated to queries and play a central role in the proposed diversification algorithms and metrics.

Recommendation diversity is also an active research topic in the Recommender Systems (RS) area. Ziegler et al. (2005) presented a topic diversification method to improve user satisfaction. Zhang et al. (2008) used a quadratic-programming approach to diversify recommendation results. Onuma et al. (2009) proposed a new recommender algorithm for producing "surprising" results. Adomavicius and Kwon (2011) address some ranking techniques to generate diverse recommendations.

3. THE RECOMMENDATION TASK

We provide first a brief reminder 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 returning a ranked list of items R in decreasing order of predicted interest. Though some personal information about the user could be used, in general the predictions are generated from the user profile (which we shall denote as $\mathbf{u}$ in boldface), i.e., those items the user has interacted with, showing some evidence of his 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.

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

Recommendations are commonly produced by computing a ranking function taking a user and an item as input. The ranking function is often a rating prediction, which we shall denote by $\hat{r}(u,i)$. We shall use $\hat{r}_{norm}(u,i)$ to denote the normalized predicted score taking values in [0,1].

4. INTENT-ORIENTED DIVERSITY IN RS

As introduced earlier, since recommendation is a particular form of retrieval task, we find it natural to consider the adaptation of diversification methods and principle that have been developed in ad-hoc IR to the RS setting. The main difference is that while search diversity principles strongly revolve around query characteristics, there is no explicit query in the recommendation task. We therefore explore potential equivalences between the two tasks that may enable a workable mapping of methods and metrics from search diversity to recommendation. First, we consider user profiles as the equivalent of search queries. Then, we need to find an analogous of query subtopics.

4.1. Use of explicit and implicit feature spaces

As an analogous of query subtopics, one might firstly consider explicit item features like, in the case of movies, their genres, directors, actors,

language, etc. The similarity between the items is conceptually easy to understand using this kind of explicit information: the more features two items share, the more similar they are. For example, using just genres, a similarity function between items could be defined with the binary cosine coefficient:

$$sim(i,j) = \frac{|\mathbf{i} \cap \mathbf{j}|}{\sqrt{|\mathbf{i}|\,|\mathbf{j}|}} \quad (1)$$

This function will be very useful for both obtaining and assessing diverse lists of recommended items.

Even more, knowing the user profile $\mathbf{u}$ it is possible to construct a user aspects space. Given a subtopic or feature f, we may estimate its probability of occurring in the aspects space of the user u as:

$$p(f|u) = \frac{|\{i \in \mathbf{u} | f \in \mathbf{i}\}|}{\sum_{f' \in \mathcal{F}} |\{i \in \mathbf{u} | f' \in \mathbf{i}\}|} \quad (2)$$

This formula will be useful for making a personalized diversification. It will be also helpful to determine an item-conditioned probability for any feature:

$$p(f|i) = \frac{[f \in \mathbf{i}]}{|\mathbf{i}|} \quad (3)$$

Another approach uses implicit information that can be extracted by the use of Matrix Factorization (MF) techniques (Koren et al. (2009)). MF extracts k-dimensional vectors of implicit features of users and items so that the inner product of vectors approximate a rating prediction, i.e., $\hat{r}(u,i) = p_u^t\, q_i$ where $p_u, q_i \in \mathbb{R}^k$. With these vectors it is possible to define a similarity function using the cosine coefficient:

$$sim(i,j) = \frac{q_i^t\, q_j}{\|q_i\|\,\|q_j\|} \quad (4)$$

It is also possible to translate this k-dimensional space in a $2k$ set of features, since each real-valued vector component may represent a preference or rejection towards a feature depending on the positiveness or negativeness of the component. This way the formulae 1, 2 and 3 can also be used in this context.

One advantage of these implicit features is that they will be always available, since they do not require any additional data apart from users interactions with the items collection. Nevertheless, since they have a too abstract meaning, we do not find them suitable for assessing diversity. Therefore they will only be used for diversification methods. The question here is whether these implicit features perform worse, similar or better with respect to the explicit ones.

4.2. IR diversification techniques to RS

We adapt two IR algorithms for search diversity to the RS re-ranking task. The first one is Maximal Marginal Relevance (Carbonell et al. (1998)), a greedy approach with the following objective function:

$$(1-\lambda)\,\hat{r}_{norm}(u,i) + \lambda\, avg_{j \in S}(dist(i,j)) \quad (5)$$

As it can be seen, the objective function represents a trade-off –parametrized by λ– between the normalized score prediction $\hat{r}_{norm}$ of the recommender and the average distance of the item to be re-ranked to the already re-ranked set S of items. The similarity function $sim(i,j)$ taking values from zero to one can be used for modelling the distance between items: $dist(i,j) = 1 - sim(i,j)$.

The other re-ranking algorithm is based on the IA-Select scheme (Agrawal et al. (2009)), it also takes a greedy approach with a different objective function:

$$\sum_{f \in \mathcal{F}} p(f|u)\hat{r}_{norm}(u,i)p(f|i) \prod_{j \in S}(1 - p(f|j)\hat{r}_{norm}(u,j)) \quad (6)$$

This function considers a sum on $f \in \mathcal{F}$ of the product of the user preference for f, the relevance of the item i for the user, and the "remaining weighted presence" of the feature in the previous re-ranked items.

5. DIVERSITY AND NOVELTY METRICS

Just as was done for the diversification techniques, it is also possible to extend IR diversity metrics to RS. The most immediate examples are α-NDCG (Clarke et al. (2008)) and the intent-aware versions of NDCG or ERR of Agrawal et al. (2009). These metrics are well defined and can consider ranking and gradual relevance, so they could be perfect candidates for assessing diversity of results recommended.

It would be also interesting to have metrics for evaluating novelty of recommendations. There are some proposals in the RS field but, as said, they do not take into account ranking and relevance. Therefore we present in the next section a new framework for defining both diversity and novelty metrics.

5.1. Item novelty and diversity

Together with similarity between items, we have identified three fundamental properties of items in a recommender system related to novelty and diversity:

- Discovery: whether an item is known or not.
- Relevance: whether an item is judged as relevant or not.
- Choice: whether an item is chosen or selected or not.

We assume that these properties can be modelled in a probabilistic way. Depending on the use of explicit or implicit preference data, there will be different estimations for the discovery and relevance of an item. We have developed natural models for global item discovery for explicit and implicit preference data, i.e., $p(seen|i)$. For more details, see Castells et al. (2011). A definition of a user-relative item discovery –$p(seen|i,u)$– does not seem so straightforward. Our current efforts focus on identifying alternative models for $p(seen|i,u)$ that take into account, for example, the knowledge of each user's community o to extract some probabilistic dependences between items, i.e., $p(i|j)$. In case we consider an item in a recommendation list, the probability of been discovered, i.e. $p(seen|i,R)$, can be simplified by a decreasing discount function $disc$.

Concerning user-relative relevance probability, i.e. $p(rel|i,u)$,we have used an estimation for rating preferences based on a utility function $g(u,i) = \max(0, r(u,i) - \tau)$ and an exponential normalized mapping as in Chapelle et al. (2009):

$$p(rel|i,u) = \frac{2^{g(u,i)} - 1}{2^{g_{max}}} \tag{7}$$

In case of implicit data we have mapped it to rating values taking and taken the same aforementioned estimation, though here other solutions may be possible. Nevertheless, we are still trying to solve the problem of sparsity of recommendations so the use of relevance does not change so abruptly the contribution of relevant and not relevant items, as you can see in the experiments section.

Finally, choice can be seen as a conjunction of discovery and relevance: $p(choose) = p(seen)\,p(rel)$.

Upon the concepts just defined we develop some item novelty models. The first one defines the novelty of an item as the probability of not having been seen (popularity complement):

$$\mathsf{PC}(i) = 1 - p(seen|i) \tag{8}$$

In case we were interested in emphasize highly novel items, a logarithmic approach can be used (inverse popularity):

$$\mathsf{IP}(i) = -\log_2 p(seen|i) \tag{9}$$

Additional one could also consider another version of the last one –assuming, for example, prior uniform probabilities among items– (free discovery):

$$\mathsf{FD}(i) = -\log_2 p(i|seen) \tag{10}$$

Another approach to the user-relative novelty modelization without $p(seen|i,u)$ uses the concept of distance between items –as seen in the previous section–. For example, one could define the novelty of an item with respect to those the user already knows as a weighted distance (profile diversity):

$$\mathsf{PD}(i|u) = \frac{\sum_{j\in\mathbf{u}} p(rel|j,u) dist(i,j)}{\sum_{j\in\mathbf{u}} p(rel|j,u)} \tag{11}$$

Finally, a similar distance-based item diversity with respect to the ranked list of recommendations is defined (intra list diversity):

$$\mathsf{ILD}(i_k|u,R) = C'_k \sum_l disc(l|k) p(rel|i_l,u) dist(i_k,i_l) \tag{12}$$

where $C'_k = 1/\sum_{l\neq k} disc(l|k)p(rel|i_l,u)$ normalizes the distance in the range $[0,1]$ and $disc(l|k) = disc(\max(1,l-k))$ is a function that considers the distance in the ranked list between elements i_l and i_k in a browsing scenario.

5.2. Resulting metrics

With the different novelty models it is possible to define a whole new set of metrics. In this subsection we define some of them and explore their connections to previously state-of-art metrics.

We use a discount model $disc(k)$ to reflect the ranking bias of a recommendation list and also the relevance probability $p(rel|i,u)$ to reflect the utility of the novelty or diversity of an item. Our metrics will have the following form:

$$C \sum_n disc(n) p(rel|i_n,u)\; f(i_n) \tag{13}$$

where $C = 1/\sum_n disc(n)$ is a normalization factor and $f(i_n)$ is one of the item novelty or diversity metrics previously described.

Starting with novelty, we define the following metrics:

- Expected popularity complement: $f_{\mathsf{EPC}} = \mathsf{PC}$
- Expected inverse popularity: $f_{\mathsf{EIP}} = \mathsf{IP}$
- Expected free discovery: $f_{\mathsf{EFD}} = \mathsf{FD}$

Note that, since IP and FD are unbounded item functions, the resulting metrics are not upper bounded.

In case we consider no discount nor relevance ($disc(n) = 1$ and $p(rel|i_n) = 1$) the EFD metric results in mean self-information (MSI), as defined in Zhou et al. (2010). In principle, EPC, EIP and EFD are designed for the same task –evaluating global novelty–, it would be interesting to determine which one of them is "better" or to find special meaningful cases where each one gives a different result.

		EFD		EPD		EILD	
	relevance	**1**	$\mathbf{p(rel\|i.u)}$	**1**	$\mathbf{p(rel\|i.u)}$	**1**	$\mathbf{p(rel\|i.u)}$
	baseline	9.8450	1.9898	0.6986	0.1420	0.6587	0.1162
	IASelect-genre	9.8601	1.8590	0.7220	0.1363	0.7183	0.1166
	IASelect-MF	9.8708	**2.0319**	0.6969	**0.1446**	0.6600	0.1187
pLSA	MMR-genre	9.9172	1.8701	**0.7666**	0.1397	**0.7899**	**0.1205**
	MMR-MF	9.9121	1.9475	0.6996	0.1388	0.6597	0.1128
	nov	**12.2518**	1.5350	0.7117	0.1080	0.6686	0.0803
	Random	11.7776	0.3165	0.7608	0.0216	0.7450	0.0129
	baseline	11.8610	1.0017	0.7572	0.0733	0.6920	0.0540
	IASelect-genre	11.3215	1.0914	0.7913	0.0819	0.8100	**0.0632**
	IASelect-MF	11.7314	**1.1482**	0.7467	**0.0832**	0.6992	0.0627
MF	MMR-genre	11.8620	0.9372	**0.8196**	0.0716	**0.8344**	0.0553
	MMR-MF	11.5011	1.0777	0.7540	0.0787	0.6952	0.0590
	nov	**16.1873**	0.1893	0.7678	0.0129	0.6217	0.0046
	Random	12.7774	0.2665	0.7702	0.0181	0.7173	0.0098
	baseline	11.5030	0.6403	0.7737	0.0475	0.7209	0.0328
	IASelect-genre	11.2379	0.8003	0.8081	0.0609	0.8259	0.0438
	IASelect-MF	11.1838	**0.9766**	0.7531	**0.0715**	0.7263	**0.0523**
kNN	MMR-genre	11.5258	0.5994	**0.8337**	0.0469	**0.8525**	0.0338
	MMR-MF	11.5743	0.6205	0.7728	0.0460	0.7239	0.0319
	nov	**13.0399**	0.1462	0.7717	0.0090	0.7055	0.0038
	Random	11.6684	0.3016	0.7713	0.0206	0.7470	0.0118

Table 1: *Results in MovieLens1M of the proposed metrics with some state-of-art recommender algorithms and re-ranking diversifiers. The best diversifications for each recommender and metric are marked in bold, and the best recommender-diversifier combination for each metric is underlined.*

With the distance-based user-relative novelty model the expected profile distance is defined:

$$\text{EPD} = C \sum_k disc(k) p(rel|i_k, u) PD(i_k|u) \qquad (14)$$

This metric will measure how different are the metrics with respect to the user profile, producing a user-relative metric for novelty.

Not so different will be the expected intra-list diversity metric, that will take the following form:

$$\text{EILD} = C \sum_k disc(k) p(rel|i_k, u) ILD(i_k|u, R) \quad (15)$$

Without considering discount or relevance EILD is equivalent to ILD seen in Ziegler et al. (2005) and Zhang et al. (2008).

6. EXPERIMENTS

The following experiments show the results the diversification methods presented measured by the proposed metrics. The experiments were done on the MovieLens1M dataset, containing one million explicit ratings from one to five for 3,900 movies by 6,040 users.

We used three different state-of-art recommenders as baselines: pLSA from Hofmann (2004), matrix factorization (MF) from Koren et al. (2009) and a k nearest neighbors (kNN). The diversification techniques previously described are used here to re-rank the 500 first results of the baseline recommenders for each user. In addition, two simple diversification techniques are also applied: a random re-ranking and a greedy diversifier that makes a trade-off between novelty and accuracy.

As shown, the results suggest that no-relevance variants give totally different results compared to the variants with $p(rel|i, u)$. While in EFD without relevance the best diversifier is always the novelty-aware one, in case of EFD with relevance the best one is IA-Select with MF-derived features. In EPD the result is similar, where best diversifier, as expected, is MMR with genre features.

Regarding the implicit features, it is interesting to see that they achieve good results for EFD and EPD, even though the metrics use movie genres as features.

The comparison between baseline recommender algorithms offers quite different results, reflecting the fact that all the three methods are of very diverse nature. With relevance metrics the best option is, by far, the pLSA algorithm; whilst in case of no relevance the best choices are MF and kNN.

No results involving rank discount reported, as we found no significant difference between adding the discount component or not.

7. CONCLUSION

The presented study proposes adaptations of IR techniques for diversifying recommendations results and a general framework for evaluating novelty and diversity of items and recommendation lists. Some specific metrics have been defined for capturing global novelty, user-relative novelty and diversity intra-list being all of them aware of important concepts such as ranking position and relevance of the items recommended.

A complete experiment on a commonly used dataset has been conducted, in which we show the effect of the diversification techniques under different metrics, and we see that the component of relevance plays a very important role in them.

As part of the future work, we have defined some main goals:

- Investigate the experimented lack of significant difference between rank-aware metrics and rank-unaware metrics.
- Assessing, through real user or statistical methods, the resulting metrics, specially the novelty variants.
- Determine whether the combination of different implicit or explicit features could provide better results for the diversification techniques.

8. REFERENCES

G. Adomavicius and Y. Kwon. 'Improving Recommendation Diversity Using Ranking-Based Techniques'. IEEE Transactions on Knowledge and Data Engineering.

R. Agrawal, S. Gollapudi, A. Halverson and S. Ieong (2009). 'Diversifying search results'. In Proceedings of the Second ACM International Conference on Web Search and Data Mining, WSDM '09, pp. 5-14, New York, NY, USA. ACM.

G. Capannini, F. M. Nardini, R. Perego and F. Silvestri (2011). 'Efficient diversification of search results using query logs'. In Proceedings of the 20th international conference companion on World wide web, WWW '11, pp. 17-18, New York, NY, USA. ACM.

J. Carbonell and J. Goldstein (1998). 'The use of MMR, diversity-based reranking for reordering documents and producing summaries'. In Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval, SIGIR '98, pp. 335-336, New York, NY, USA. ACM.

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

O. Chapelle, D. Metlzer, Y. Zhang and P. Grinspan (2009). 'Expected reciprocal rank for graded relevance'. In Proceeding of the 18th ACM conference on Information and knowledge management, CIKM '09, pp. 621-630, New York, NY, USA. ACM.

C. L. A. Clarke, M. Kolla, G. V. Cormack, O. Vechtomova, A. Ashkan, S. Büttcher and I. MacKinnon (2008). 'Novelty and diversity in information retrieval evaluation'. In Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval, SIGIR '08, pp. 659-666, New York, NY, USA. ACM.

T. Hofmann (2004). 'Latent semantic models for collaborative filtering'. ACM Transactions on Information Systems 22(1):89-115.

Y. Koren, R. Bell and C. Volinsky (2009). 'Matrix Factorization Techniques for Recommender Systems'. Computer 42(8):30-37.

Q. Mei, J. Guo and D. Radev (2010). 'DivRank: the interplay of prestige and diversity in information networks'. In Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining, KDD '10, pp. 1009-1018, New York, NY, USA. ACM.

K. Onuma, H. Tong and C. Faloutsos (2009). 'TANGENT: a novel, 'Surprise me', recommendation algorithm'. In KDD '09: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 657-666, New York, NY, USA. ACM.

R. L. T. Santos, et al. (2010). 'Exploiting query reformulations for web search result diversification'. In Proceedings of the 19th international conference on World wide web, WWW '10, pp. 881-890, New York, NY, USA. ACM.

M. Zhang and N. Hurley (2008). 'Avoiding monotony: improving the diversity of recommendation lists'. In Proceedings of the 2008 ACM conference on Recommender systems, RecSys '08, pp. 123-130, New York, NY, USA. ACM.

T. Zhou, et al. (2010). 'Solving the apparent diversity-accuracy dilemma of recommender systems'. Proceedings of the National Academy of Sciences 107(10):4511-4515.

C.-N. Ziegler, et al. (2005). 'Improving recommendation lists through topic diversification'. In Proceedings of the 14th international conference on World Wide Web, WWW '05, pp. 22-32, New York, NY, USA. ACM.