An analysis on Time- and Session-aware diversification in recommender systems

Vito W. Anelli, Vito Bellini, Tommaso Di Noia,
Wanda La Bruna,Paolo Tomeo, Eugenio Di Sciascio*
Polytechnic University of Bari
Via E. Orabona, 4
Bari, Italy
{vitowalter.anelli,vito.bellini,tommaso.dinoia,wanda.labruna,paolo.tomeo,eugenio.disciascio}@poliba.it

ABSTRACT

In modern recommender systems, diversity has been widely acknowledged as an important factor to improve user experience and, more recently, intent-aware approaches to diversification have been proposed to provide the user with a list of recommendations covering different aspects of her behavior. In this paper, we propose and analyze the performances of two diversification methods taking into account temporal aspects of the user profile: in the first one we adopt a temporal decay function to emphasize the importance of more recent items in the user profile while in the second one we perform an evaluation based on the identification and analysis of temporal sessions. The two proposed methods have been implemented as temporal variants of the well-known xQuAD framework. In both cases, experimental results on Netflix 100M show an improvement in terms of accuracy-diversity balance .

ACM Reference format:
Vito W. Anelli, Vito Bellini, Tommaso Di Noia,
Wanda La Bruna,Paolo Tomeo, Eugenio Di Sciascio. 2016. An analysis on Time- and Session-aware diversification in recommender systems. In *Proceedings of 25th Conference on User Modeling, Adaptation and Personalization, Bratislava, Slovakia, 9-12th July (UMAP'17),* 5 pages.
DOI: http://dx.doi.org/xx.xxxx/xxxxxxx.xxxxxxx

1 INTRODUCTION

Recommender systems are designed to meet the users' needs suggesting relevant items in a personalized fashion. As recommendations are usually presented in form of list or group, the user experience strongly depends on the overall quality of such recommendations and, the diversity among them has been identified as one of the most important quality factor [6, 13]. Generally, accuracy and diversity are considered as contrasting properties, due to the demonstrated trade-off between them in offline evaluation [6]. In spite of that, a recent user's behaviour study proved that diversity in recommendations has an important positive impact on user satisfaction [8]. Moreover, in [15] the authors show that by taking into account the users propensity towards diversity, it is possible to foster the recommendation diversity without affecting accuracy or even slightly improve it. The diversity issue has been originally addressed in the Information Retrieval field. As user queries are often ambiguous and their intent is not clear, proposing a set of answers covering different intents may increase the probability that users find at least one relevant document [7, 17]. The concept of intent-aware diversification has then been applied to the Recommender Systems field [21] and extensively studied thus producing new algorithms and evaluation metrics [18, 20, 22]. Here, user intents as defined in Information Retrieval have been mapped to user interests with reference to item characteristics.

Very often, in the design of the model behind a recommendation engine, the user profile is considered as a static snapshot without taking into proper account its temporal dimension. Actually, the importance of analyzing temporal aspect for user modeling has been proved to affect the final recommendation results [9–11].

Based on the above observations, in this paper we investigate the effect on the trade-off between accuracy and diversity of a recommendation list when dealing with temporal aspects of the user profile. The intuition behind our idea is that temporal dynamics might allow to better understand the user interests with respect to the items characteristics and then provide a more accurate intent-aware diversification. Therefore, this work presents two intent-based modeling methods that exploit the time dimension in a user profile. The first one analyses the frequency of interaction between the users and the items features using a temporal decay function in order to emphasize *persistence* and *recency* of an intent. The other method is based of a new session analysis technique of user ratings for intent modeling. Considering that a session is usually defined as a set of consecutive ratings with a very small gap of time among them (e.g. less than one hour in music [10]), we provided a wide definition of session tailored for movie ratings. In particular, such method is designed to highlight *importance*, *persistence* and *recency* of an intent among user sessions. We experimentally evaluated such methods with the large scale movie dataset published in the context of Netflix Prize Context [4]. The experimental results demonstrated that the analysis of temporal aspects in the user profile leads to better accuracy-diversity balance and intent-aware diversity compared to the original xQuAD. As an additional benefit, the aggregate diversity results improved too thus demonstrating to produce more personalized recommendations.

To the best of our knowledge, this is the first attempt in the investigation of how temporal dynamics affect diversity in recommendations.

*Authors are listed in alphabetical order.

UMAP'17, Bratislava, Slovakia

DOI: http://dx.doi.org/xx.xxxx/xxxxxxx.xxxxxxx

2 RELATED WORK

A list of recommendations can be diversified in an implicit or explicit manner [3]. While the implicit diversification is used to increase the average distance between pairs of items in the recommendation list, the explicit method diversifies the recommendations trying to cover the user interests represented via categories or other descriptive information of the items. Therefore, the explicit diversification is known as Intent-Aware since it considers the likeness of user intents in information retrieval and user interests in recommender systems [21]. Explicit Query Aspect Diversification (xQuAD) is one of the most well-known intent-aware framework originally proposed for query results diversification [17] and then adapted to the recommendation [20] field. In a nutshell, xQuAD aims at maximizing the coverage of the inferred interests while minimizing their redundancy in a recommendation list.

The importance of taking into account the temporal dynamics in recommender systems has been recently pointed out in different works for diverse recommendation domains. A method to model user sessions in music domain was proposed in [10]. It considers as session each set of consecutive ratings without an extended time gap between them. Considering that there are various psychological phenomena that lead to a set of ratings to be grouped into a single session, such method captures these effects by means of user session biases. [11] presented a collaborative filtering algorithm able to model time drifting of user preferences and the results on the Netflix dataset indicated the importance of unveiling temporal effects in order to produce more accurate recommendations. A more recent method proposed to take advantage of temporal information in user behavior is called Time-based Markov Embedding [9], used to find the best next-song recommendation via Latent Markov Embedding.

In this work we aim at exploring the exploitation of temporal dynamics in user intents to provide a better intent-aware diversification.

3 INTENT-AWARE DIVERSIFICATION FOR RECOMMENDATIONS

Typically, a recommender system produces a list of personalized recommendations for each user. According to [1], a re-ranking of such list can be applied to improve its diversity, without modifying the recommendation process. However, finding the most diverse results is a NP-hard problem and hence several heuristics have been proposed [12]. Most previous diversification approaches are based on a greedy selection strategy [2, 5, 17]. Such strategy selects the next most relevant item only if that item is diverse with respect to the items already selected [12]. Algorithm 1 describes the working scheme of a greedy selection method. For the purpose of this work, we consider xQuAD, one of the most well-known intent-aware greedy heuristics. It maximizes the coverage of the inferred interests while minimizing their redundancy. xQuAD was proposed for search diversification in information retrieval by Santos et al. [17], as a probabilistic framework to explicitly model an ambiguous query as a set of sub-queries that are supposed to cover the potential aspects of the initial query. More recently, it has been adapted for recommendation diversification by Vargas and Castells [20], replacing query and relative aspects with user and items features,

Data: The original list $\overline{\mathrm{R}}$, $N \leq n$
Result: The re-ranked list $\overline{\mathrm{S}}$

$\overline{\mathrm{S}} = \langle\rangle$;
while $|\mathrm{S}| < N$ **do**
  $i^* = \underset{i \in \mathrm{R} \setminus \mathrm{S}}{\mathrm{argmax}}\, f_{obj}(i, \overline{\mathrm{S}}, u)$;
  $\overline{\mathrm{S}} = \overline{\mathrm{S}} \circ i^*$;
  $\mathrm{R} = \mathrm{R} \setminus \{i^*\}$
end
return $\overline{\mathrm{S}}$.

Algorithm 1: The greedy strategy.

respectively. The expression of the xQuAD objective function is

$$f_{obj}(i, \overline{\mathrm{S}}, u) = \lambda \cdot r^*(u, i) + (1 - \lambda) \cdot div(i, \overline{\mathrm{S}}, u) \tag{1}$$

with $div(i, \overline{\mathrm{S}}, u)$ defined as

$$div(i, \overline{\mathrm{S}}, u) = \sum_f p(i|f) \cdot p(f|u) \cdot \prod_{j \in \mathrm{S}} (1 - p(j|f)) \tag{2}$$

In (2) $p(i|f)$ represents the likelihood of item i being chosen given the feature f and is computed as a binary function that returns 1 if the item contains f, 0 otherwise; $p(f|u)$ represents the interest of user u in the feature f and is usually computed as the relative frequency of the feature f on the items rated by user u. In other words, xQuAD fosters the idea of promoting items that are simultaneously highly related to at least one of the features of interest for the user and slightly related to the features of the items already recommended. In particular, this work focuses on the intent modeling in the xQuAD framework, namely the aforementioned $p(f|u)$ component in the Equation (2).

4 INTENT MODELING WITH TEMPORAL DYNAMICS

In this section, we propose two methods to exploit temporal analysis for intent modeling in diversification that we call **session-based** and **time-based intent modeling**. Both relies on the intuition that user intent can change during the interaction with the system and evaluating the importance of a feature merely computing its frequency in the user profile may not represent the current user interests.

4.1 Time-Based Intent Modeling

In order to valorize *persistence* and *recency* of an intent, we propose to analyze the frequency of interaction between the user u and the feature f and to weight each interaction by a temporal decay function. More formally, the following formula computes the interest of the user u with respect to the feature f:

$$p(f|u) = \frac{\sum_{i \in R(u)} cov(f, i) disc(u, i)}{\sum_{i \in R(u)} disc(u, i)} \tag{3}$$

where $R(u)$ indicates the set of rating provided by the user u; $cov(f, i)$ is a binary function returning 1 if the item i is associated with the feature f, 0 otherwise; $disc(i, u)$ is a temporal decay function returning lower values for older ratings, and higher values for the most recent ones. Inspired by [11], as decay function we adopted the following exponential function

$$disc(u,i) = e^{-\beta \cdot |t_{u,last} - t_{u,i}|} \tag{4}$$

where $t_{u,last}$ indicates the timestamp of the last rating of the user u and $t_{u,i}$ the timestamp when user u rated i; $\beta > 0$ controls the decay rate.

In our experimental setting we adopted the Netflix dataset which contains ratings from October 1998 until October 2005. Hence, a value of β equals 1/128 would bring to a not considerable weight of $2 * 10^{-9}$. We decided to take into account all the ratings even if they are quite far away in time with a small but yet considerable weight of $2,831 * 10^{-6}$ using a β value of 1/200.

4.2 Session-Based Intent Modeling

User sessions definition. Session analysis is quite common in music domain, since users are used to listen to many songs in sequence. There a session is represented by a set of consecutive ratings with a small gap of time between them [10]. Conversely, sessions are not easy to find in movie domain, since users typically watch a small number of movies in brief timeslots and the temporal gap among visions or ratings could be large (sometimes several days or even months). In our setting, in order to identify user sessions we propose an EM clustering used to train two univariate Gaussian Mixture Models (with equivariance and variable variance). The number of clusters has been evaluated based on the Bayesian Information Criterion considering the fitted models with a number of clusters from 1 up to 300. In order to remove outliers from each session s, for each computed cluster we do not consider ratings falling outside the interval $[\mu_s - \sigma_s, \mu_s + \sigma_s]$, with μ_s and σ_s being respectively the mean and the standard deviation of ratings distribution for the session s.

Intent modeling. Once user sessions are determined, they can be used to analyze the user activities taking into account the temporal dynamics. In this work we present an approach to model the users intents over time, by considering three key properties: *importance*, *persistence* and *recency* of an intent among the user sessions. The first property indicates the importance of an intent in each session computed as the percentage of items covering that intent. The second property considers how many sessions the intent is important for, therefore it sums the importance of the intent for each sessions. Finally, the third property focuses on the intent freshness, penalizing old sessions with a temporal decay function.

More formally, the following formula computes the interest of the user u with respect to the feature f:

$$p(f|u) = \frac{\sum_{s \in S(u)} \frac{\sum_{i \in I(s)} cov(f,i)}{|F(s)|} disc(s,u)}{\sum_{s \in S(u)} disc(s,u)} \tag{5}$$

where $S(u)$ indicates sessions computed for the user u; $I(s)$ is the set of items in s; $cov(f,i)$ is a binary function returning 1 if the item i is associated with the feature f, 0 otherwise; $F(s)$ represents the set of features associated with all the items in s; $disc(s)$ is the temporal decay function adapted to handle the sessions instead of the items, considering a session as an item in Equation (4) where the session date is that of the last rated item in such session. As for the previous case, β value was set to 1/200.

5 EXPERIMENTAL SETTING

Dataset. In order to verify our research questions and evaluate our proposal, we used the popular movie datasets derived from the Netflix Prize Context [4]. Netflix dataset contains over 100 million ratings provided by ˜480,000 users on ˜17,000 movies. Such ratings were collected between 1998 and 2005 and associated with the relative date. However, such dataset contains noise added on purpose for reasons of privacy, as explained in the Netflix Prize Rules*: "some of the rating data for some customers in the training and qualifying sets have been deliberately perturbed in one or more of the following ways: deleting ratings; inserting alternative ratings and dates; and modifying rating dates". Indeed, we found that some users rated an exaggerated number of movies in some days: 30% of all the users have rated at least 61 movies in the most *prolific* day. Therefore, in order to train the session-based user model on a clean subset of the dataset, we selected a sample of users removing the outliers by means of the following steps: (i) we discarded the users with less than 20 ratings as at this stage of our study we are not interested in evaluating the cold users behavior; (ii) we ordered the users in decreasing order of the maximum number of daily interactions and discarded the top 30%; (iii) we ordered the users in decreasing order of the average number of daily interactions and discarded the top 30%. The dataset used for training the models contains 233,452 users, 18,104,476 rating and 17,763 movies. We built training and test sets by employing a 80%-20% holdout temporal split for each user.

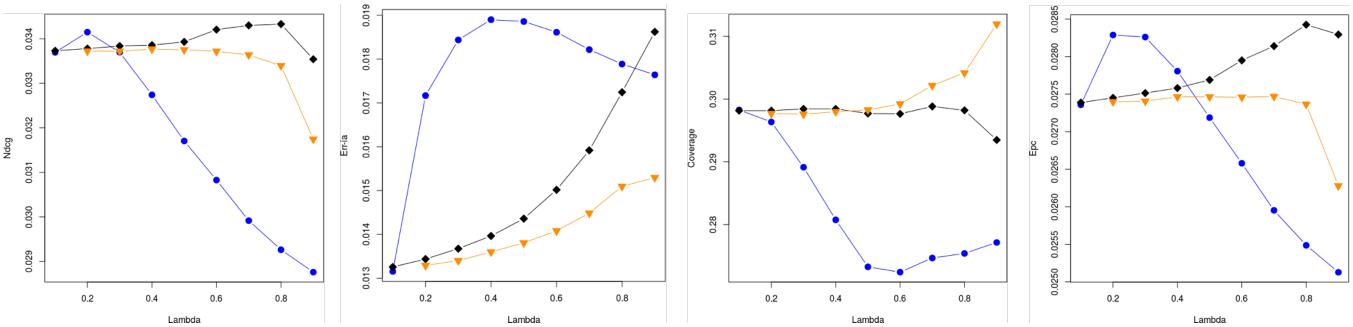
Recommendation algorithms. We evaluated our approaches w.r.t. the xQuAD baseline via a re-ranking of the BPRMF [16] and BPRSLIM [14] algorithms resulting recommendations. We trained both models using the MyMediaLite† implementation upon the dataset described in Section 5 to produce for each user a Top-300 recommendations list ($\overline{\mathbf{R}}$ in Algorithm 1, used to compute $r^*(u,i)$ in Equation (1)), and then we re-ranked those lists using xQuAD. The resulting recommendations lists are the baselines we compare against. xQuAD uses side information to lead to diversified recommendations. In this work the diversification is based on the movie genre feature. In Netflix dataset this information is not explicitly provided. In order to extract the genre information, we mapped each movie with the corresponding Freebase resource by means of its title and year of release and we then selected the corresponding genres. Overall, the number of distinct genres extracted is 266.

In our evaluation, the time-based and session-based intent modelings proposed in Section 4 are used as alternatives to the pure frequency based intent modeling in the original xQuAD. These two variations of xQuAD, are denoted as: TB_xQuAD, SB_xQuAD, where *TB* stands for time-based and *SB* for session-based. The resulting evaluated algorithms are then:

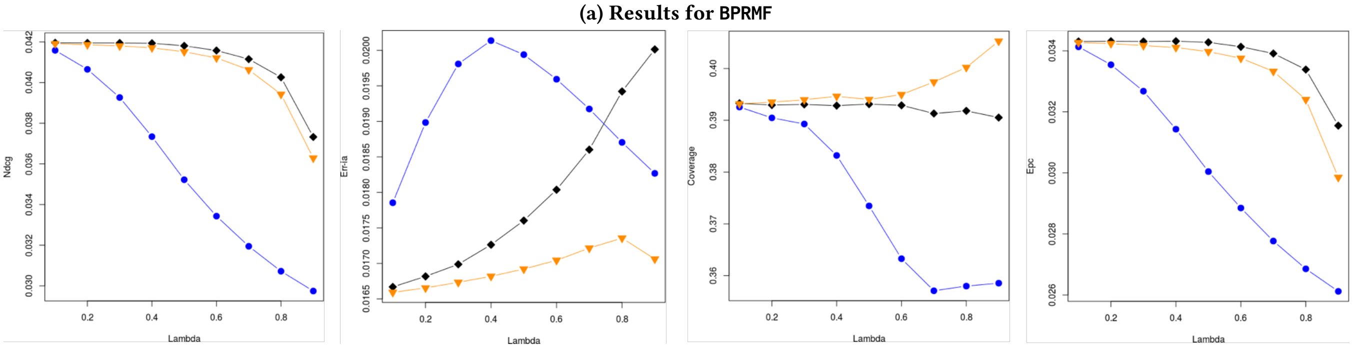
- BPRMF + xQuAD (baseline)
- BPRMF + Time-based xQuAD variant (TB_xQuAD)
- BPRMF + Session-based xQuAD variant (SB_xQuAD)
- BPRSLIM + xQuAD (baseline)
- BPRSLIM + Time-based xQuAD variant (TB_xQuAD)
- BPRSLIM + Session-based xQuAD variant (SB_xQuAD)

*http://netflixprize.com/rules.html

†http://www.mymedialite.net



(a) Results for BPRMF



(b) Results for BPRSLIM

Figure 1: Curves obtained by varying λ from 0.1 to 0.9 (step 0.1), using BPRMF and BPRSLIM as recommendation algorithm. From left to right we plot the values of nDCG, ERR-IA, Catalog Coverage and EPC. The blue line represents values for the base xQuAD evaluation while the black and yellow lines represent respectively values for the time-based and session-based version of xQuAD

Metrics. In order to evaluate *accuracy*, we measured nDCG. As for *individual diversity*, namely the degree of dissimilarity among all items in the list provided to a user, was measured by ERR-IA as it has been shown [19] to be the metric targeted by xQuAD, while for *aggregate diversity* we computed the Catalog Coverage (percentage of items recommended at least to one user). An evaluation on the *novelty* of computed results has been done through EPC (Expected Popularity Complement) [19]. For all the aforementioned metrics we used the implementation provided by RankSys framework[‡] on the Top-10 recommendation list.

Results Discussion. Charts in Figure 1a and Figure 1b show the curves for nDCG, ERR-IA, Catalog Coverage and EPC, for both BPRMF and BPRSLIM variants. Very interestingly both the time- and session-based version of xQuAD improves results in terms of accuracy, aggregate diversity as well as of novelty independently of the recommendation algorithm adopted. Generally, the time-based variant of xQuAD performs better than the session-based one but for Catalog Coverage where we have better results for the session-based implementation of xQuAD. It is worth noticing that the base version of xQuAD outperforms its time-based variants up to a certain value of λ for both BPRMF and BPRSLIM. For the former this value lies between 0.8 and 0.9 while for the latter between 0.7 and 0.8. Hence, in case we are interested in higher values of diversity, time may play an important role. This observation is also strengthen by the higher values we obtain in terms of precision, catalog coverage and novelty. In Table 1 we see that with $\lambda = 0.8$ we obtain the best result in terms of trade-off among the various metrics we measured in our experiments.

Algorithm	Ndcg@10	ERR IA@10	Coverage@10	EPC@10
BPRMF+XQUAD	0.029264	**0.01789**	0.27540	0.02549
BPRMF_SB_XQuAD	0.03340	0.01510	**0.30417**	0.02737
BPRMF_TB_XQuAD	**0.03433**	0.01724	0.29820	**0.02843**
BPRSLIM+XQUAD	0.03072	0.01870	0.35799	0.02686
BPRSLIM_SB_XQuAD	0.03943	0.01736	**0.40021**	0.03240
BPRSLIM_TB_XQuAD	**0.04026**	**0.01942**	0.39183	**0.03339**

Table 1: Comparative results in terms of accuracy, individual diversity and aggregate diversity with $\lambda = 0.8$

6 CONCLUSIONS AND FUTURE WORK

In this paper we investigate the role of temporal information while modeling a user profile in computing diversified recommendations. We propose two different time-dependent user modelings which take into account also the user rating history. One of the two proposed methods bases on a new session analysis technique by considering those periods where the user interacted in a more constant way with the system. Experimental results demonstrated that considering temporal dynamics leads to better accuracy-diversity balance and better intent-aware diversification. The results we obtained in this preliminary investigation are quite promising and we are currently extending our experimental evaluation to different datasets in diverse domains as well as to other metrics in order to measure the quality of the recommendations not just in terms of accuracy when time is taken into account.

[‡]https://github.com/RankSys/RankSys

REFERENCES

[1] Gediminas Adomavicius and YoungOk Kwon. 2012. Improving Aggregate Recommendation Diversity Using Ranking-Based Techniques. *IEEE Transactions on Knowledge and Data Engineering* 24, 5 (May 2012), 896–911. DOI: http://dx.doi.org/10.1109/TKDE.2011.15

[2] Azin Ashkan, Branislav Kveton, Shlomo Berkovsky, and Zheng Wen. 2015. Optimal Greedy Diversity for Recommendation.. In *IJCAI*, Qiang Yang and Michael Wooldridge (Eds.). AAAI Press, 1742–1748.

[3] Fabiano Belém, Rodrygo Santos, Jussara Almeida, and Marcos Gonçalves. 2013. Topic Diversity in Tag Recommendation. In *Proceedings of the 7th ACM Conference on Recommender Systems (RecSys '13)*. ACM, New York, NY, USA, 141–148. DOI: http://dx.doi.org/10.1145/2507157.2507184

[4] James Bennett, Stan Lanning, and Netflix Netflix. 2007. The Netflix Prize. In *In KDD Cup and Workshop in conjunction with KDD*.

[5] Jaime Carbonell and Jade 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)*.

[6] Pablo Castells, Neil J. Hurley, and Saul Vargas. 2015. *Novelty and Diversity in Recommender Systems*. Springer US, Boston, MA, 881–918. DOI: http://dx.doi.org/10.1007/978-1-4899-7637-6_26

[7] Olivier Chapelle, Shihao Ji, Ciya Liao, Emre Velipasaoglu, Larry Lai, and Su-Lin Wu. 2011. Intent-based Diversification of Web Search Results: Metrics and Algorithms. *Inf. Retr.* 14, 6 (Dec. 2011), 572–592. DOI: http://dx.doi.org/10.1007/s10791-011-9167-7

[8] Michael D. Ekstrand, F. Maxwell Harper, Martijn C. Willemsen, and Joseph A. Konstan. 2014. User Perception of Differences in Recommender Algorithms. In *Proceedings of the 8th ACM Conference on Recommender Systems (RecSys '14)*. ACM, New York, NY, USA, 161–168. DOI: http://dx.doi.org/10.1145/2645710.2645737

[9] Ke Ji, Runyuan Sun, Wenhao Shu, and Xiang Li. 2015. Next-song Recommendation with Temporal Dynamics. *Know.-Based Syst.* 88, C (Nov. 2015), 134–143. DOI: http://dx.doi.org/10.1016/j.knosys.2015.07.039

[10] Noam Koenigstein, Gideon Dror, and Yehuda Koren. 2011. Yahoo! Music Recommendations: Modeling Music Ratings with Temporal Dynamics and Item Taxonomy. In *Proceedings of the Fifth ACM Conference on Recommender Systems (RecSys '11)*. ACM, New York, NY, USA, 165–172. DOI: http://dx.doi.org/10.1145/2043932.2043964

[11] Yehuda Koren. 2009. Collaborative Filtering with Temporal Dynamics. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '09)*. ACM, New York, NY, USA, 447–456. DOI: http://dx.doi.org/10.1145/1557019.1557072

[12] Onur Küçüktunç, Erik Saule, Kamer Kaya, and Ümit V. Çatalyürek. 2014. Diversifying Citation Recommendations. *ACM Transactions on Intelligent Systems and Technology* 5, 4, Article 55 (Dec. 2014), 21 pages. DOI: http://dx.doi.org/10.1145/2668106

[13] Sean M. McNee, John Riedl, and Joseph A. Konstan. 2006. Being Accurate is Not Enough: How Accuracy Metrics Have Hurt Recommender Systems. In *CHI '06 Extended Abstracts on Human Factors in Computing Systems (CHI EA '06)*. 1097–1101.

[14] Xia Ning and George Karypis. 2011. Slim: Sparse linear methods for top-n recommender systems. In *Data Mining (ICDM), 2011 IEEE 11th International Conference on*. IEEE, 497–506.

[15] Tommaso Di Noia, Jessica Rosati, Paolo Tomeo, and Eugenio Di Sciascio. 2017. Adaptive multi-attribute diversity for recommender systems. *Inf. Sci.* 382-383 (2017), 234–253.

[16] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. 2009. BPR: Bayesian personalized ranking from implicit feedback. In *Proceedings of the twenty-fifth conference on uncertainty in artificial intelligence*. AUAI Press, 452–461.

[17] R.L.T. Santos, C. Macdonald, and I. Ounis. 2010. Exploiting Query Reformulations for Web Search Result Diversification. In *WWW '10*. 881–890.

[18] Saúl Vargas, Linas Baltrunas, Alexandros Karatzoglou, and Pablo Castells. 2014. Coverage, Redundancy and Size-awareness in Genre Diversity for Recommender Systems. In *Proceedings of the 8th ACM Conference on Recommender Systems (RecSys '14)*. 209–216.

[19] Saúl Vargas and Pablo Castells. 2011. Rank and Relevance in Novelty and Diversity Metrics for Recommender Systems. In *Proceedings of the Fifth ACM Conference on Recommender Systems (RecSys '11)*. 109–116.

[20] S. Vargas and P. Castells. 2013. Exploiting the Diversity of User Preferences for Recommendation. In *OAIR '13*. 129–136.

[21] Saul Vargas, Pablo Castells, and David Vallet. 2011. Intent-oriented Diversity in Recommender Systems. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '11)*. ACM, New York, NY, USA, 1211–1212. DOI: http://dx.doi.org/10.1145/2009916.2010124

[22] Jacek Wasilewski and Neil Hurley. 2016. Intent-Aware Diversification Using a Constrained PLSA. In *Proceedings of the 10th ACM Conference on Recommender Systems (RecSys '16)*. ACM, New York, NY, USA, 39–42. DOI: http://dx.doi.org/10.1145/2959100.2959177