Coverage, Redundancy and Size-Awareness in Genre Diversity for Recommender Systems

Saúl Vargas[a]*, Linas Baltrunas[b], Alexandros Karatzoglou[b], Pablo Castells[a]
[a]Universidad Autónoma de Madrid, Spain
[b]Telefónica Research, Spain
saul.vargas@uam.es, linas@tid.es, alexk@tid.es, pablo.castells@uam.es

ABSTRACT

There is increasing awareness in the Recommender Systems field that diversity is a key property that enhances the usefulness of recommendations. Genre information can serve as a means to measure and enhance the diversity of recommendations and is readily available in domains such as movies, music or books. In this work we propose a new Binomial framework for defining genre diversity in recommender systems that takes into account three key properties: genre *coverage*, genre *redundancy* and recommendation list *size-awareness*. We show that methods previously proposed for measuring and enhancing recommendation diversity –including those adapted from search result diversification– fail to address adequately these three properties. We also propose an efficient greedy optimization technique to optimize Binomial diversity. Experiments with the Netflix dataset show the properties of our framework and comparison with state of the art methods.

Categories and subject descriptors: H3.3 [Information Search & Retrieval]: Information Filtering

Keywords: Recommender Systems; Diversity; Genres

1. INTRODUCTION

Recommender Systems [1] are intelligent personalized Information Retrieval tools where the information need of a user is fully or partially expressed by means of her profile or history rather than a query. The Recommender Systems literature has mostly focused on optimizing the accuracy of their results, either by predicting the preference for an item by a user (rating prediction task) or by selecting a list of items to present to the user (top-N recommendation task). Focusing solely on accuracy involves the risk of producing dull recommendations that do not capture all the facets of interest to the users. Additional properties such as diversity, novelty, explicability and context-awareness are key to expand the users' options and make recommendations more informative and useful. This paper tackles the diversity problem for recommendations.

Most users have quite diverse tastes even within the same domain such as movies or books. For example, in the movie domain the same user may like Action but also Drama movies. Standard recommender systems, especially content based algorithms [1], fail to address this diversity of tastes. Recommendation *list diversification* techniques can solve the user's need for more varied recommendations and help her discover new products (music, movies, apps). Diversity is a list-wise property that has been shown to enhance the user satisfaction with respect to the recommendations [14]. Several notions of diversity have been proposed in the field and, although they are closely related, they are not equivalent. In this paper we focus on the notion of intra-list diversity [22], i.e., providing a list of varied recommendations that covers the different interests of the user. We define the notion of diversity using genres, which are used in domains such as books, movies and music.

We analyze the properties of genres and their utility in providing diverse recommendations. We postulate three important properties that genre-based diverse recommendations should fulfill: 1) *genre coverage*, that is, each genre should be represented in a recommendation list according to both the interest of the user and its specificity; 2) *redundancy*: while it is important that all genres are represented it is equally important not to over-represent a particular genre – this is particularly important in domains where items can have more than one genre; and 3) recommendation *list size-awareness*, which focuses on the common screen space limitation to offer recommendations, and how it influences genre coverage and redundancy. Our analysis of state of the art diversification methods and metrics shows that they do not properly or fully address these three properties. We propose a new Binomial framework that takes into account all the aforementioned properties. The framework consists of a metric to assess the diversity of recommendations and a greedy re-ranking strategy to optimize the diversity of recommendations. We report experiments on a widely-known dataset for recommendation – Netflix[1] – showing the properties of our framework, and comparing it to state of the art methods.

2. RELATED WORK

We start by reviewing the related work and positioning our research with respect to the state-of-the-art. First, we present diversity as a key dimension of recommendation utility, and compare it to notions of diversity developed in the field of Information Retrieval. Second, we present the current state-of-the-art techniques for modeling recommendation diversity that our work compares to.

2.1 Diversity in Recommender Systems

Along with the progress targeting accuracy in Recommender Systems, researchers have realized that improving

*Work conducted during research stay at Telefónica Research.

RecSys'14, October 6–10, 2014, Foster City, Silicon Valley, CA, USA.

ACM 978-1-4503-2668-1/14/10 ...$15.00.
http://dx.doi.org/10.1145/2645710.2645743.

[1]http://www.netflixprize.com/

recommendations' usefulness and user satisfaction may require more than being accurate. In particular, Herlocker et al. [10] stated that accuracy alone may not give users of recommender systems an effective and satisfying experience. McNee et al. [13] further specified that there are properties other than accuracy that have an effect on user satisfaction, namely coverage, diversity, novelty or serendipity.

Diversity in Recommender Systems, that is, addressing the user's varied tastes and his/her need for diverse recommendations, has been shown to help improve the attractiveness and usefulness of recommendations [14]. In this paper we focus on the so-called intra-list distance as defined by Ziegler et al. [22], i.e., how different are the items in a recommendation list with respect to each other. In [22], Ziegler et al. propose taxonomies of products that are used to define a similarity metric between items, although other sources of diversity could be considered. For instance, in our previous work [19] we used movie genres as the source of diversity. On the other hand, Kabutoya et al. [12] and Shi et al. [17] extract latent topic models from the users' interactions with the system in order to create diverse recommendations.

The problem of intra-list diversity in Recommender Systems is related to search result diversification [7]. Users of general-purpose commercial search engines tend to submit short queries to represent their information needs. Such brevity in queries tends to lead to ambiguity – the query could have many possible interpretations – and underspecification, – the topic the query refers to may have different facets. A way to cope with this problem is the diversification of web search results i.e., presenting lists of documents that cover as many interpretations or facets as possible of the original query as early as possible in the ranking [15]. The term *subtopic* is a commonly used word in this area for referring to query interpretations or facets altogether.

2.2 Measuring and enhancing diversity

Different frameworks for measuring and enhancing the diversity of recommendation lists have been proposed in the Recommender Systems and Information Retrieval literature. We briefly recall here the most closely related and relevant research to the scope of our work. Based on different principles, most of them propose re-ranking an initial recommendation list so that diverse items can be shown early in the list. The most common approach is based on greedy selection, which tends to achieve a good and efficient approximation to an optimal re-ranking. Methods of this type fit in the algorithmic structure described in Algorithm 1. The algorithm depends on the specific definition of an objective function f_{obj} which defines the marginal utility of an item with respect to items ranked above it.

Algorithm 1 A greedy selection of the items in recommendation list R to produce a re-ranked list S.

```
S = ∅
while |R| > 0 do
    i* = arg max_{i∈R\S} f_obj(i; S)
    R = R \ {i*}
    S = S ∪ {i*}
end while
return S
```

One of the earliest and best-known proposals for diversity in Recommender Systems is the "topic list diversification" from Ziegler et al.[22], to which we will refer as **pair-wise framework** throughout this paper. This framework defines a diversity metric called intra-list similarity (ILS), as the sum of similarities between all pairs of items in the recommendation:

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

where sim is a similarity measure between items. In previous work [18], we defined an extension to this metric known as expected intra-list diversity (EILD), which allows to consider the relevance and position of the recommended items. Ziegler et al. also proposed a greedy re-ranking strategy to optimize ILS which is structurally equivalent to the maximal marginal relevance (MMR) by Carbonell and Goldstein [5]:

$$f_{MMR}(i;S) = (1-\lambda)\, rel(i) + \lambda \min_{j \in S} dist(i,j) \tag{2}$$

where λ is a trade-off parameter between the original ranking and the diversity component and $rel(i)$ is the relevance of the item i. Zhang and Hurley [21] also considered the problem of optimizing ILS as a quadratic optimization problem.

Another major line of work in measuring and enhancing diversity comes from search result diversification. In particular, the **intent-aware framework** [2] considers the sum of the weighted marginal relevance of each subtopic s of a query, as it is the case of the intent-aware version of the ERR metric [6]:

$$ERR-IA = \sum_s p(s) \sum_{k=1}^{|R|} \frac{1}{k} rel(i_k) \prod_{l=1}^{i-1} (1 - rel(i_l)) \tag{3}$$

where $p(s)$ is the probability of s being the intended subtopic behind the query. Santos et al. [15] proposed the *explicit query aspect diversification* (xQuAD), a re-ranking approach that optimizes intent-aware metrics by enhancing the coverage of the different subtopics while minimizing their redundancy:

$$\begin{aligned} f_{xQuAD}(i;S) = {} & (1-\lambda)\, p(i) \\ & + \lambda \sum_s p(s)\, p(i|s) \prod_{j \in S} (1 - p(j|s)) \end{aligned} \tag{4}$$

where $p(i|s)$ is the probability of choosing item i given the subtopic s. The IA-Select re-ranking approach of Agrawal et al. [2], with minor differences, can be considered as a subcase of xQuAD with $\lambda = 1.0$. This intent-aware framework was adapted to Recommender Systems in [19], by translating the concept of query subtopic to user aspects.

A third approach is the more recent **proportionality framework** by Dang and Croft [9] for search result diversification. They emphasize the need for covering each subtopic of the search query by offering a number of relevant documents proportional to the interest of the subtopic they cover. The basis for measuring this proportionality is the so-called disproportionality metric, defined as:

$$DP = \sum_s \mathbf{1}_{v_s \geq k_s^R} (v_s - k_s^R)^2 + \frac{1}{2} n_{NR}^2 \tag{5}$$

where v_s is the expected number of documents that cover the subtopic s, k_s^R the actual number of documents, and n_{NR} the number of non relevant documents. On top of DP, Dang and Croft propose a cumulative proportionality metric (CPR) that is the basis of their study. Analogously to the other proposals they define a greedy re-ranking approach, the proportionality method (PM), inspired on a seat assignment system for legislative elections in some countries:

$$\begin{aligned} f_{PM}(i;S) = {} & \lambda \frac{v_{s*}}{1 + 2\sum_{j \in S} \frac{p(j|s^*)}{\sum_{s'} p(j|s')}}\, p(i|s^*) \\ & + (1-\lambda) \sum_{s \neq s^*} \frac{v_s}{1 + 2\sum_{j \in S} \frac{p(j|s)}{\sum_{s'} p(j|s')}}\, p(i|s) \end{aligned} \tag{6}$$

where s^* indicates the least-covered subtopic in S. Note that this proportionality framework admits a straightforward adaptation to recommendation similar to that of the intent-aware framework.

3. CHARACTERIZING GENRES

As defined in the Merriam-Webster dictionary[2], a genre is "a category of artistic, musical, or literary composition characterized by a particular style, form, or content". We argue that genres can be used as the source for defining diversity as they:

- explicitly define a conventional style of an item that has a common interpretation among users,
- have the potential of representing the different tastes of individual users,
- are well accepted for media categorization and are already available in most online media catalogs for movies, literature, music, etc.
- and it is safe to assume that the user will perceive the diversity of the recommendation list if the genres are diversified among the recommended items. Other alternatives such as using item-to-item distance based on consumption patterns may have an effect on the inherent diversity of the recommendation, although this may not directly translate to a user perception of diversity.

Genres, nonetheless, present some particularities that need to be addressed to be used effectively. First, genres can have different levels of generality: for example, in the movie domain "Drama" represents a very broad and vaguely defined style with many diverse movies belonging to this genre. On the other hand, "Western" is a quite specific movie type which is usually devoted to telling stories in the American Wild West. This generality is also reflected in the number of items for each genre. See Table 1 for the number of movies in each genre in the Netflix data set. We observe that the generality of each genre is also related to the perception of redundancy in a recommendation list. For example, three random westerns in a short recommendation list of five items feels more redundant than three random dramas. We will exploit this observation when defining our probabilistic model.

Genre	Count	Genre	Count	Genre	Count
Action	1,464	Fantasy	651	Romance	1,887
Adult	54	Film-Noir	70	Sci-Fi	819
Adventure	996	Game-Show	2	Short	237
Animation	381	History	317	Sport	284
Biography	384	Horror	900	Talk-Show	2
Comedy	3,025	Music	568	Thriller	1,989
Crime	1,319	Musical	418	War	422
Documentary	779	Mystery	709	Western	285
Drama	4,408	News	1		
Family	772	Reality-TV	15		

Table 1: Genre distribution in Netflix.

Second, genres do not usually define disjoint or isolated categories in their domains, and it is generally difficult to establish a precise hierarchy among them. For example, "The Lord of the Rings" by Tolkien can be classified as Adventure, Fiction, High fantasy and British literature all at once. Moreover, careless use of sub-genres can lead to lower perceived diversity. For example, heavy metal and white metal – two closely related sub-genres – share the same musical techniques, modes of dress and performance and could be perceived as similar by a listener.

In order to illustrate the aforementioned properties of genres, we present the case of movie recommendations of the Netflix dataset. This dataset contains one hundred million ratings (from 1 to 5 stars) from 480,000 to 17,770 different movies. Using IMDb[3], we found genre information for 9,320 movies in the dataset accounting for 83% of the ratings in the dataset.

[2]`http://www.merriam-webster.com`

[3]`http://www.imdb.com/`

As seen in Table 1 the number of movies for each genre varies greatly, from 4,408 movies in "Drama" to only 1 movie in "News". Genres do not form disjoint categories, as seen in Figure 1, which shows the overlap between the top 5 genres by a Venn diagram. One can see that, for instance, there are only 76 pure "Romance" movies, and the other 96% of movies in this genre overlap with at least one other genre. Other genres also have a high degree of overlap. In fact, there is no clear hierarchical structure between the genres. It also seems that overlaps between genres do not follow any particular distribution. Furthermore, pairwise overlaps between genres are not wide enough as to establish any clear sub-genre relationship between one another; even the narrowest and most specific genres (for example, Crime) have only partial overlaps (<60%) with more general genres such as Drama.

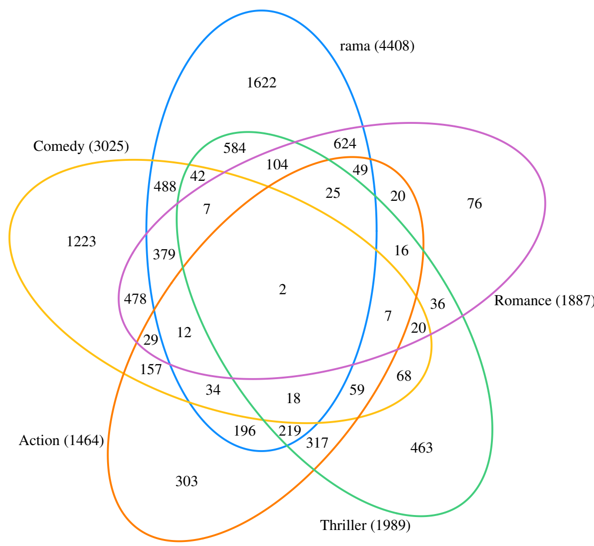

Figure 1: Venn diagram for the 5 most frequent genres in the Netflix dataset.

4. MEASURING GENRE DIVERSITY IN RECOMMENDATION LISTS

We all have an intuitive idea of what genre diversity means for a list of movies. Yet when it comes to translating the intuition to a mathematical expression that reflects degrees of diversity by a numeric value, one has to be more specific about what the value should reflect. In particular, drawing from the IR diversity literature [2, 5, 20], two different dimensions should be considered to this respect, namely genre coverage and redundancy. We take them as required properties that a genre-based recommendation diversity metric should capture. Furthermore, we argue that these dimensions should be captured in a way that takes into account the properties of genres discussed and exemplified in Section 3. Moreover, we add to these a third and new requirement, size-awareness, which has not been explicitly considered in prior work. We briefly discuss each of these three properties next.

Coverage is the simplest and most obvious property. Since most users enjoy items from a variety of genres, it is important that the recommendation list covers as many of them as possible. Moreover, this coverage should be proportional: even when a user is interested in several genres, the personalized importance of each genre is not equal. Therefore, the more a user is interested in a given genre, the more important it is that it is covered in the recommendation list.

Second, **redundancy** should also be considered. It is not enough to have a high coverage of genres in order to have a diverse recommendation list. We may put it this way: it is as important to present items that cover a certain genre as to present other items that do *not* cover it. This notion of redundancy should take into account the preferences for the user as well as how general each genre is. Consider the extreme example shown in Table 2 where three movies are recommended to a user. Even if these 3 movies cover a total of 6 genres, the diversity is not quite perceivable. This is because all three movies cover a very narrow Western genre which makes the recommendation list highly redundant.

Finally, **size-awareness**. Coverage and redundancy should depend on the length of the recommendation list. Since the rise of mobile devices, the issue of having limited screen real estate to show recommendations requires a careful selection of what to display in that list. We also improve over existing diversity enhancing techniques by specifically addressing the recommendation list size. For example, when generating a short recommendation list one should only recommend items from the most relevant genres. In a longer list we could have higher genre redundancy depending on the generality of the involved genres. To the best of our knowledge this kind of adaptation has not been explored in prior work on search or recommendation diversity.

The reviewed techniques in Section 2.2 do not satisfy all these properties, in particular:

- The intra-list similarity of Ziegler et al. [22] is defined as a pairwise property of elements in a list. A pair-wise property does not translate however as directly as we may expect to a list-wise property as we are stating. Further, it is not trivial to consider a similarity measure that takes into account by itself the generality of different genres and the user-specific importance of each of them. Essentially equivalent to the approach by Ziegler et al, the MMR scheme [5] displays the same limitations.
- The intent-aware framework (IA-metrics, IA-select and xQuAD) [2, 15] considers coverage and redundancy, but as to the latter, the scheme does not fully capture the view that it is equally important to present items that cover a certain genre as to present other items that do not cover it. Specifically, the redundancy component of ERR-IA and xQuAD reduces the contribution of items that cover redundant genres, rather than discounting them as negative from the list diversity value. Thus, items covering a redundant genre will contribute positively to the diversity even though the contribution diminishes with each additional occurrence of the genre. Furthermore, this redundancy does not detract at all from the contribution of additional genres the items can have in addition to the redundant one – that is, the genres are assumed to be totally independent from each other. The example in Table 2 illustrates this effect: it is fine (diversity-wise) in the context of this framework that all the movies in the recommendation list be westerns, as long as they cover also other genres. As a consequence, the diversifications are biased to retrieve items that cover many genres. We may reasonably question the implicit assumption in this scheme that multiple genres in the same item will procure the same diversity perception as multiple genres over different items.
- The work by Dang and Croft [9] does cover an idea of user-centric proportionality, but redundancy is not penalized and therefore, it may also suffer from the same problems as xQuAD for genre diversity.
- None of the prior search or recommendation diversification methods takes into account the size of the retrieved list that will be presented to (or browsed by) the user. The diversification schemes have therefore no means to consider this information to enhance diversity at a particular rank cutoff.

Movie	Genres
Wild Wild West	Action, Comedy, Sci-Fi, Western
Cowboys and Aliens	Action, Sci-Fi, Thriller, Western
The Good, the Bad and the Ugly	Adventure, Western

Table 2: Example of redundant recommendations

5. A BINOMIAL FRAMEWORK FOR GENRE DIVERSITY

A naïve approach for creating diverse recommendations consists in making a random selection of items. This approach offers highly diverse recommendations, but it tends to approximate the poorest possible output in terms of the relevance of recommendations for the user interests, which makes it an option of little practical use. Still, the nature of the selection of genres in a random recommendation provides a meaningful basis to build a revised notion of diversity upon it. In particular, we propose to use a binomial distribution to model how a personalized recommendation would match a random recommendation in terms of the diversity of genres, using the binomial distribution to model the likelihood that a given genre will appear by chance in a recommendation, and take this as a reference to assess the diversity value of a given genre distribution among recommended items. In essence, this approach means considering random item recommendation as the optimal approach in terms of pure genre diversity, and using a binomial distribution as the model for the genre distribution resulting from random item sampling.

5.1 The Binomial Diversity Metric

The binomial distribution is the discrete probability distribution of the number k of successes in a sequence of N independent Bernoulli trials with the same probability of success p. A random variable that follows this distribution, $X \sim B(N, p)$, has the following probability mass function:

$$P(X = k) = \binom{N}{k} p^k (1-p)^{N-k} \quad (7)$$

We base our definition of a genre diversity metric on top of this as follows. For each genre, we measure its coverage and redundancy using binomial distributions. We consider the selection of an item covering each genre as a Bernoulli trial, whereby for each genre, a recommendation list can be viewed as a sequence of Bernoulli trials. It must be noticed that these trials are not independent: a recommendation list is actually a selection without replacement. However, given that the typical recommendation list size is usually much smaller than the set of movies covering each genre, we can treat these trials as if they were independent, and therefore use the binomial distribution to model how likely is a genre to appear in a recommendation list.

More formally, for an item i and a set of genres $G(i)$ covered by the item i, we consider the Bernoulli experiment of whether a randomly sampled genre g belongs to $G(i)$. Given a set of items S, we denote the number of items belonging to that genre – the number of successes – as $k_g^S = |\{i \in S \ : \ g \in G(i)\}|$. Given a recommendation list R of size N, we take the probability of a genre p_g as a measure of how "adequate" is the number k_g^R of items covering a genre g in that recommendation. As required in Section 4, this probability should take into account the generality of a genre and also the relevance of each genre for the user. We

	i_1	i_2		ILD	S-recall	ERR-IA	CPR	Coverage	NonRed	BinomDiv
				Yes	Yes	**Not Always**	**Not Always**	Yes	Yes	Yes
Better	b	a	Regardless of	1.0000	0.6666	**0.4000**	**0.7857**	0.8255	1.0000	0.8255
Worse	a	a	p(a), p(b)	0.0000	0.3333	**0.5000**	**0.8571**	0.6814	0.3333	0.2269
				No(=)	Yes	Yes	Yes	Yes	**No(=)**	Yes
Better	a c	b	Regardless of	**1.0000**	1.0000	0.7000	0.9643	1.0000	**1.0000**	1.0000
Worse	a	b	p(a), p(b), p(c)	**1.0000**	0.6666	0.5000	0.8929	0.8255	**1.0000**	0.8255
				Yes	**No(=)**	**No(<)**	**No(<)**	**No(=)**	Yes	Yes
Better	a	b	Regardless of	1.0000	**0.6666**	**0.5000**	**0.8929**	**0.8255**	1.0000	0.8255
Worse	a b	b	p(a), p(b)	0.5000	**0.6666**	**0.6500**	**0.9286**	**0.8255**	0.3780	0.3120
				No(=)	**No(=)**	Yes	Yes	Yes	Yes	Yes
Better	a	a	p(a) > p(b)	**0.0000**	**0.3333**	0.5000	0.8571	0.6814	0.3333	0.2271
Worse	b	b		**0.0000**	**0.3333**	0.2500	0.6429	0.5200	0.1429	0.0743

Table 3: Postulates of diversity. Each of the four postulates shows two rankings (displayed horizontally) with better or worse diversity. Each item in the ranking is represented by the genres (a, b, c) that it belongs to. We show in the table the diversity score that each of the metric assigns to the lists. In the computation of the metric values, we assume for simplicity there are only three genres in the dataset, with prior probabilities, as an example, p(a) = 0.5, p(b) = 0.25, p(c) = 0.25. For ERR-IA we use the same definition as in the TREC diversity task (as computed by the ndeval script), generalized to support non-uniform aspect distributions.

propose to combine global genre distribution statistics and personalized user preferences to estimate p_g as follows.

On one side, the relevance of a genre for the user u can be estimated by using historical data, i.e., considering the local proportion p''_g of the items the user has had some interaction with, denoted as $\mathcal{I}_u$. On the other side, the generality of the genre can be estimated by the global proportion p'_g of items in the user preferences covering it. To join both global and local probabilities, we propose a simple linear combination:

$$p''_g = \frac{k_g^{\mathcal{I}_u}}{|\mathcal{I}_u|} \quad p'_g = \frac{\sum_u k_g^{\mathcal{I}_u}}{\sum_u |\mathcal{I}_u|} \quad p_g = (1-\alpha)\, p'_g + \alpha\, p''_g \qquad (8)$$

With all the components of genre-based binomial distributions, we now define scores for the coverage and redundancy of a recommendation list R. We measure coverage as a property defined by the genres that are present in the recommendation list and those that are not. The maximum coverage would be achieved when all the genres of interest are covered in the recommendation list. However, this maximum is not always reachable, especially in small recommendation lists. Therefore, when some genres cannot be covered, the coverage should reflect the loss caused by their absence, which should be proportional to their importance. We thus define the coverage score as the product for the genres not represented in the recommendation list of their probabilities of not being randomly selected according to X_g, normalized by the $|G|$-th root:

$$Coverage(R) = \prod_{g \notin G(R)} P(X_g = 0)^{1/|G|} \qquad (9)$$

We define redundancy, in turn, only by the genres covered in the recommendation list. The moment one genre appears more than once in a recommendation list, it can be potentially redundant, although not all genres will be equally affected. We model the redundancy of a genre appearing k times in a recommendation list by a "remaining tolerance" score that reflects how probable it would be that the genre appeared at least k times in a random list:

$$P(X_g \geq k \mid X_g > 0) = 1 - \sum_{l=1}^{k-1} P(X_g = l \mid X_g > 0) \qquad (10)$$

Some examples of this "remaining tolerance" score are illustrated in Figure 2. The non-redundancy score is consequently defined as the product of the "remaining tolerance" scores for each covered genre, normalized by the $|G(R)|$-th root:

$$NonRed(R) = \prod_{g \in G(R)} P(X_g \geq k_g^R \mid X_g > 0)^{1/|G(R)|} \qquad (11)$$

Finally, the Binomial Diversity metric is defined as the product of both components:

$$BinomDiv(R) = Coverage(R) \cdot NonRed(R) \qquad (12)$$

The previous definition can be adapted to consider only the relevant recommended items by re-defining k_g^R as the number of relevant items covering the genre g and the number of trials N as the number of relevant recommended items.

Note that binomial relevance satisfies all the properties described in Section 4. It maximizes the coverage of the genres according to their p_g. It takes into account user preferences via p''_g. It penalizes over-represented genres by rapidly decreasing their redundancy score. Lastly, it is adapted to the recommendation length by parameter N.

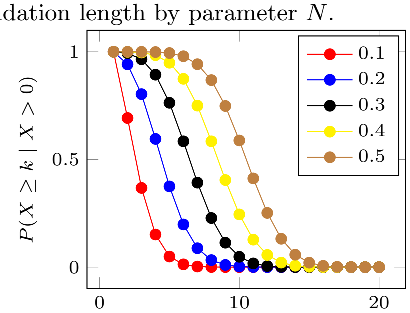

Figure 2: $P(X \geq k \mid X > 0)$ for different values of p and k of binomial distributions with $N = 20$ (continuous lines are drawn just as a reference).

5.2 A Binomial Re-ranking Algorithm

A greedy re-ranking approach to optimize binomial diversity can be straightforwardly derived from the proposed metric scheme by just defining an objective function that linearly combines relevance and binomial diversity as follows:

$$f_{BinomDiv}(i; S) = (1-\lambda)\, rel(i) + \lambda\, div(i; S) \qquad (13)$$

where the relevance component $rel(i)$ can be defined as the score that the baseline recommender system assigns to the items for the target user, and the diversity component is the difference in terms of the binomial diversity in Equation 12 after adding the candidate item to the re-ranked recommendation list:

$$div(i; S) = BinomDiv(S \cup \{i\}) - BinomDiv(S) \qquad (14)$$

Because we are combining variables with different ranges and distributions, in the practical implementation of the objective function we need to normalize both scores. In our experiments we do so by transforming them to z-scores, that is, we subtract their mean and divide by the standard deviation: $norm_X(x) = \frac{x - \mu_X}{\sigma_X}$.

5.3 Qualitative analysis

In addition to the empirical behavior of the proposed scheme, the Binomial Diversity metric fulfills qualitative properties that further specify the requirements stated earlier in Section 4. These properties can be formalized by four postulates shown in Table 3, which we propose as a basis on which diversification metrics can be analyzed and compared to each other, providing a clear way to show properties of each metric, identify and report the differences, in a similar perspective as proposed in [3]. Each postulate presents a rule, which expresses a simple idea on how we can reason about the genre-based diversity. We represent each of the postulates by providing two ranked lists of items (displayed horizontally in the table) with minimal differences. The ranked list denoted by "Better" should have strictly higher diversity that the one denoted by "Worse". For example, the first postulate expresses the idea that a ranked list of two items that cover two genres (a and b) is more diverse than a list of two items that cover only one genre (a). We mark a method with "Yes" only if the metric complies with the postulate, otherwise we indicate to what extent the metric fails to satisfy the postulated inequality (either the metric yields the opposite inequality, or is insensitive to the difference between the two lists). We can see that all of the state of the art methods fail at least one of the tests, and only our proposed Binomial Diversity that combines Coverage and NonRed properties complies with all the postulates. For illustration, we show in the same table the diversity score that each of the analyzed diversification metric assigns to the prototypical lists.

In order to further illustrate how the diversification metric works and to show the benefits of the genre-based approach, we may examine the working example shown in Table 4. The example shows the top 20 recommended movies by the item-based kNN method ($R_0 \cup R_1$) for a sample user from the Netflix dataset, and the re-ranking of this list by the binomial diversification ($R_0 \cup R_2$), shown by the movies that are removed (R_1) and added (R_2) as a result of the re-ranking. The first row of the table summarizes the user taste profile (p''_g), i.e. what fraction of movies of each genre he has rated. We see that the user is inclined towards Drama, Comedy and Action movies. We may also notice that the user seldom watched War movies. Both recommendation lists have an overlap of 11 movies (R_0) that are shown below the user profile information. If we compare the differences between both recommendation lists – the kNN baseline $R_0 \cup R_1$ and its diversification by the binomial scheme $R_0 \cup R_2$, we see that the baseline promotes Action and War movies that are over-represented in the final list of 20 movies, thus creating a highly redundant recommendation. The recommender under-represents other genres such as Comedy which plays a major part in the user profile. The binomial diversification, on the other hand, uses the p''_g and the list size as the reference for how many movies of each genre it should select to avoid redundancy. There are already 7 Action movies in the list and, therefore, it promotes several Comedies instead. Moreover, it includes new genres such as Animation, Children's and Mystery that help improve the coverage score. This leads to a significant increase of the diversification score for the diversified list.

6. EXPERIMENTS

In order to show the properties of the Binomial Diversity framework, we have carried out two experiments on two common datasets for movie recommendation: the MovieLens1M[4] collection and the data from the Netflix Prize. In

[4]http://movielens.umn.edu/

	Movie	Action	Adventure	Animation	Children's	Comedy	Crime	Drama	Mystery	Romance	Sci-Fi	Thriller	War	Western
	$\mathbf{p''_g}$	0.25	0.20	0.04	0.08	0.39	0.13	0.42	0.07	0.19	0.15	0.25	0.06	0.02
Kept (R_0)	Braveheart	X						X					X	
	Jerry Maguire							X		X				
	Matrix, The	X									X	X		
	Negotiator, The	X										X		
	Patriot Games	X										X		
	Pulp Fiction						X	X						
	The Silence of the Lambs							X				X		
	Terminator 2	X									X	X		
	Titanic							X		X				
	Total Recall	X	X								X	X		
	True Lies	X	X			X				X				
Removed (R_1)	Air Force One	X										X		
	Enemy of the State	X										X		
	Get Shorty	X				X		X						
	Gladiator	X						X						
	Green Mile, The							X				X		
	Independence Day	X									X		X	
	Schindler's List							X					X	
	Star Wars: Episode V	X	X					X			X		X	
	Star Wars: Episode VI	X	X							X	X		X	
Added (R_2)	As Good As It Gets					X		X						
	Back to the Future III					X					X			X
	Elizabeth							X						
	Erin Brockovich							X						
	The Game								X			X		
	Leon: The Professional						X	X		X		X		
	South Park			X		X								
	There's Sth. About Mary					X								
	Toy Story			X	X	X								

Table 4: Binomial diversification in action.

the case of the Netflix Prize, no genre information was provided in the data, so we extracted that information from IMDb. Discarding the movies for which we could not find genres, the resulting dataset includes 83 million ratings (on a 1-5 scale) by 480,000 users for 9,320 movies classified into 28 different genres. Due to the similarities between the results for both datasets and the space limit, we only report and discuss the results for the Netflix dataset.

6.1 Setup

We split the rating data into training and test in a 5-fold cross validation. We followed a common evaluation procedure [4, 8] in which for each target user the recommenders are required to rank a list which includes both relevant and non-relevant items. The relevant items include all those having a test rating for the target user, where the rating value is above a threshold. In the set of irrelevant items we include all movies with a test rating value below the threshold, plus a set of 1000 randomly sampled movies.

We take for our experiment two baseline collaborative filtering (CF) algorithms: an item-based nearest neighbors recommender [16] (Item-kNN) and the implicit Matrix Factorization algorithm (iMF) by Hu et al. [11]. We additionally include two non-personalized systems for further reference: recommendation by item popularity (PopRec) and random recommendation (Random). On each of the CF baselines, we applied the diversification approaches described in Section 2.2 (MMR, PM and xQuAD) and our Binomial diversification. The optimal value of the λ parameter in these diversifiers is set by a grid search in the $[0, 1]$ interval by steps of 0.1. We additionally report as a reference the effects of random re-ranking.

We evaluate the diversity of the recommendations with respect to genres by our proposed Binomial Diversity metric (BinomDiv), plus the diversity metrics presented in Section 2.2: EILD, ERR-IA and CPR. We report two addi-

	nDCG	BinomDiv			CPR			EILD	S-recall
		$\alpha = \mathbf{0.0}$	$\alpha = \mathbf{0.5}$	$\alpha = \mathbf{1.0}$	$\alpha = \mathbf{0.0}$	$\alpha = \mathbf{0.5}$	$\alpha = \mathbf{1.0}$		
Random	0.0172	**0.4391**	**0.4286**	0.2834	**0.7948**	0.7550	0.6857	**0.8147**	**0.6022**
PopRec	0.2988	0.2297	0.2886	0.2600	0.7905	0.7960	0.7632	0.7781	0.5087
Item-kNN	0.3762	0.2232	0.3035	0.2926	0.7847	**0.8154**	0.8006	0.7753	0.5125
iMF	**0.5221**	0.2091	0.3090	**0.3257**	0.7631	0.8121	**0.8131**	0.7488	0.5175

Table 5: Results for cut-off 20 of the recommendation baselines applied to the Netflix dataset. The results for the metrics BinomDiv, CPR, EILD and S-recall consider here the diversity provided by all recommended items. Results in bold indicate the best result for a given dataset and metric.

	nDCG	BinomDiv		CPR		EILD		ERR-IA	S-recall		SPI
		all	rel	all	rel	all	rel		all	rel	
iMF	**0.5221**	0.3090	0.2501	0.8121	0.5188	0.7488	0.2318	0.2176	0.5175	0.2933	2.7851
Random	0.1502	0.3683	0.1975	0.8235	0.2233	0.7740	0.2342	0.0724	0.5579	0.1758	2.7641
BinomDiv (0.7)	0.3740	**0.6890**	**0.3191**	**0.9398**	0.4367	0.8007	0.2465	0.1664	0.6817	0.2850	2.5899
PM (1.0)	0.4671	0.3035	0.2593	0.8863	0.5179	0.7423	0.2292	*0.2194*	0.5232	0.2889	2.9931
MMR (0.9)	0.3468	0.5996	0.3044	0.7278	0.3528	**0.8898**	**0.2716**	0.1866	**0.7387**	0.2618	2.3329
xQuAD (0.2)	0.5050	0.2837	0.2421	0.8504	**0.5252**	0.7446	0.2305	**0.2347**	0.5390	**0.3060**	**3.0295**

Table 6: Results for cut-off 20 and $\alpha = 0.5$ for different recommendation algorithms and their diversified re-rankings applied to the Netflix dataset. For diversifications, the parameter chosen (in parenthesis) is the one that achieves the best result with respect to its relevance-aware objective metric. Bold indicates the best result for a given baseline and metric. Italics indicate non statistically significant differences to the corresponding baseline (Wilcoxon $p < 0.05$).

tional simple diversity metrics: subtopic recall [20] (S-recall), which is the total number of returned genres (as a ratio on the total number of genres in the collection) in a recommendation; and the average number of subtopics (genres) per item (SPI), which will serve as a reference to study the bias discussed in Section 4 for the intent-aware framework. We consider two variants for the BinomDiv, EILD, CPR and S-recall metrics: one, denoted as *all*, where the genres of all recommended items are taken into account to measure the diversity of the list, and another, denoted as *rel*, in which only those items relevant to the user are taken into account. Complementarily to the diversity metrics, we report the normalized discount cumulative gain (nDCG) to check the relevance-based effectiveness (regardless of diversity) of all configurations.

6.2 Results for Baseline Diversity

Table 5 shows the results for all the recommender baselines without any diversification step. All the metrics are evaluated at a 20 ranking cut-off. The algorithms are sorted by their nDCG score. We omit in this table the *rel* variants since they strongly correlate with nDCG when comparing recommendation algorithms with very different levels of accuracy, and thus they are not informative. For the metrics BinomDiv and CPR we show three alternatives (see Equation 8): global or non-personalized ($\alpha = 0.0$), intermediate ($\alpha = 0.5$) and fully personalized ($\alpha = 1.0$).

As we can see, the random recommender, as expected, has a very low accuracy, but scores very high for all diversity metrics, especially the non-personalized ones (BinomDiv and CPR with $\alpha = 0.0$, EILD and S-recall). The popularity-based recommendation has a much higher accuracy than the random recommendations, but has generally lower scores for diversity metrics, specially in BinomDiv, where it is in general the worst alternative. The personalized recommenders, which have a higher accuracy than the non-personalized recommenders, tend to score low in terms of non-personalized diversity metrics (BinomDiv and CPR with $\alpha = 0.0$, EILD and S-recall), but clearly improve in BinomDiv and CPR when the user history is considered ($\alpha > 0.0$).

The results from Table 5 show consistent results, showing that random recommendations are diverse in nature, and personalized recommendations may benefit from a re-ranking diversification step.

6.3 Results for Diversified Results

We present in Table 6 the results of diversifying the iMF recommendation baseline. Diversifications of the other personalized recommendation (Item-kNN) were also carried out with similar outcomes, but we omit them because of the space limit. All the metrics are computed again at a 20 ranking cut-off. The α parameter in BinomDiv (metric and diversifier) and CPR is set to 0.5. As mentioned before, the λ parameter value (shown in parenthesis) for the binomial, PM, MMR and xQuAD diversifications is the one that maximizes the corresponding objective metric: BinomDiv-rel, CPR-rel, EILD-rel and ERR-IA, respectively.

A first overall trend we may observe is that, in terms of nDCG, all the diversifications involve a decrease in the accuracy of the recommendations, showing an also expectable trade-off between relevance and diversity. Second, each diversifier is always the best option with respect to its target metric, with the exception of PM, which is outperformed by BinomDiv in CPR-all and by xQuAD in CPR-rel. This can be explained because PM does not optimize directly the formulation of CPR (see Equations 5 and 7), while the rest of the diversifiers optimize directly their target metric. Third, for CPR-rel the improvements over the baseline are almost imperceptible, and restricted to xQuAD. This shows that a diversification algorithm (and the metric it is intended to target) devised for a search task may not get the expected results in a recommendation setting with a different subtopic-document (in our case, genre-item) distribution patterns, which is one of the motivations for our framework.

As to the intent-aware framework –the xQuAD diversifier and the ERR-IA metric– and the proportionality framework –the PM diversifier and the CPR metric–, the results evidence one of the problems pointed out in Section 4, namely the accumulation of genres without any penalization for redundancy. The SPI values show how the xQuAD and PM strategies notably increase the average number of genres per item in the diversified recommendations, which, as discussed in Section 4, does not necessarily fit well with an effective notion of diversity in recommendation. Regarding the metrics, ERR-IA and CPR-rel show a clear correlation with SPI, a bias that narrows the informativeness of this metric in a recommendation setting.

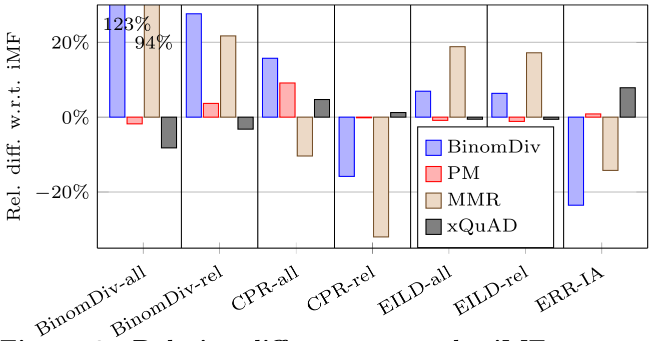

Figure 3: Relative difference over the iMF recommender baseline of the binomial, PM, MMR and xQuAD diversifiers on Netflix.

	all			rel		
	5	**10**	**20**	**5**	**10**	**20**
iMF	0.4282	0.3533	0.3090	0.4374	0.3429	0.2501
5	**0.9110**	0.5712	0.3984	**0.5692**	0.3753	0.2571
10	0.8535	**0.8906**	0.5523	0.5412	**0.4506**	0.2705
20	0.7811	0.8483	**0.8656**	0.5169	0.4229	**0.3191**

Table 7: Results at cutoffs $N = 5, 10, 20$ for the Binomial Diversity metric and diversifier ($\alpha = 0.5$) on Netflix. Bold indicates the best diversification cut-off for a each metric cut-off in each dataset.

Another key observation in the diversity-specific analysis concerns the relation between the frameworks. To facilitate the analysis of the results of Table 6, we show in Figure 3 the relative improvements over the baseline. The cross-wise relationships between diversifications and metrics from different frameworks show interesting findings. Overall, we can see in the figure that the binomial diversifier improves over the baseline when measured with BinomDiv, CPR-all and EILD, showing that the binomial framework is able to promote the proportionality of CPR-all and the dissimilarity of items of EILD by improving the coverage and non-redundancy of the recommendations. In turn, the proportionality of PM does not seem sufficient to promote the diversity when measured with BinomDiv, while the dissimilarity of MMR improves the coverage and non-redundancy of BinomDiv. The intent-aware framework shows some relation with the proportionality framework, most probably caused by the aforementioned accumulation of genres without penalization of redundancy, and does not offer improvements in the other frameworks. All these observations support our postulation that the binomial framework is able to capture and procure coverage while avoiding redundancy, uncovering a diversity angle beyond what the other frameworks can capture.

Finally, in order to evaluate the size-awareness of our Binomial framework, Table 7 shows the correspondence, using the iMF baseline, between the cut-off of the binomial diversification algorithm (the N in 13) and the cut-off of the binomial diversity metric. For each diversification cut-off, the results correspond to the best λ of the objective function. As expected, the best diversification cut-off always agrees with the cut-off of the diversity metric, in both *all* and *rel* variants. This shows that our approach is able to leverage knowledge of the desired result set size in order to bring an additional made-to-fit improvement at the targeted cut-off, a feature that is not supported in any prior framework.

7. CONCLUSIONS

We tackle in this paper the problem of diversity using genre information in Recommender Systems. An analysis of the properties of genres helps us define the requirements that a genre-based definition of diversity in recommendation should satisfy, namely coverage, non-redundancy and recommendation list size-awareness. We propose a binomial framework that satisfies these properties. A metric is defined upon this framework, and a greedy re-ranking algorithm that optimizes it. Experiments on two movie recommendation datasets validate the consistency of our framework, illustrate its properties, and show they comply with the stated requirements. As future work, we will extend our experiments to other datasets, such as music and books, and carry out user studies to further analyze and contrast the behavior and properties of the proposed framework.

8. ACKNOWLEDGMENTS

This work was supported by the national Spanish project TIN2013-47090-C3-2 and the European Union's Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 610594 (CrowdRec).

9. REFERENCES

[1] G. Adomavicius and A. Tuzhilin. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE TKDE*, 17(6):734–749, June 2005.

[2] R. Agrawal, S. Gollapudi, A. Halverson, and S. Ieong. Diversifying search results. *WSDM* 2009, pp. 5–14.

[3] E. Amigó, J. Gonzalo, and F. Verdejo. A General Evaluation Measure for Document Organization Tasks. *SIGIR* 2013, pp. 643–652.

[4] A. Bellogin, P. Castells, and I. Cantador. Precision-oriented evaluation of recommender systems: an algorithmic comparison. *RecSys* 2011, pp. 333–336.

[5] J. Carbonell and J. Goldstein. The use of MMR, diversity-based reranking for reordering documents and producing summaries. *SIGIR* 1998, pp. 335–336.

[6] O. Chapelle, D. Metlzer, Y. Zhang, and P. Grinspan. Expected reciprocal rank for graded relevance. *CIKM* 2009, pp. 621–630.

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

[8] P. Cremonesi, Y. Koren, and R. Turrin. Performance of recommender algorithms on top-n recommendation tasks. *RecSys* 2010, pp. 39–46.

[9] V. Dang and W. B. Croft. Diversity by proportionality: an election-based approach to search result diversification. *SIGIR* 2012, pp. 65–74.

[10] J. L. Herlocker, J. A. Konstan, L. G. Terveen, and J. T. Riedl. Evaluating collaborative filtering recommender systems. *ACM ToIS*, 22(1):5–53, Jan. 2004.

[11] Y. Hu, Y. Koren, and C. Volinsky. Collaborative filtering for implicit feedback datasets. *ICDM* 2008, pp. 263–272.

[12] Y. Kabutoya, T. Iwata, H. Toda, and H. Kitagawa. A probabilistic model for diversifying recommendation lists. *WTA*, volume 7808 of *LNCS*, pp. 348–359, 2013.

[13] S. M. McNee, J. Riedl, and J. A. Konstan. Being accurate is not enough: how accuracy metrics have hurt recommender systems. *CHI* 2006, pp. 1097–1101.

[14] P. Pu, L. Chen, and R. Hu. A user-centric evaluation framework for recommender systems. *RecSys* 2011, pp. 157–164.

[15] R. L. Santos, C. Macdonald, and I. Ounis. Exploiting query reformulations for web search result diversification. *WWW* 2010, pp. 881–890.

[16] B. Sarwar, G. Karypis, J. Konstan, and J. Riedl. Item-based collaborative filtering recommendation algorithms. *WWW* 2001, pp. 285–295.

[17] Y. Shi, X. Zhao, J. Wang, M. Larson, and A. Hanjalic. Adaptive diversification of recommendation results via latent factor portfolio. *SIGIR* 2012, pp. 175–184.

[18] S. Vargas and P. Castells. Rank and relevance in novelty and diversity metrics for recommender systems. *RecSys* 2011, pp. 109–116.

[19] S. Vargas, P. Castells, and D. Vallet. Explicit relevance models in intent-oriented information retrieval diversification. *SIGIR* 2012, pp. 75–84.

[20] C. X. Zhai, W. W. Cohen, and J. Lafferty. Beyond independent relevance: methods and evaluation metrics for subtopic retrieval. *SIGIR* 2003, pp. 10–17.

[21] M. Zhang and N. Hurley. Avoiding monotony: improving the diversity of recommendation lists. *RecSys* 2008, pp. 123–130.

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