UCL DEPARTMENT OF
COMPUTER SCIENCE

Research Note
RN/11/21

Auralist: Introducing Serendipity into Music Recommendation

12 December 2011

Yuan Cao Zhang, Diarmuid Ó Séaghdha, Daniele Quercia, Tamas Jambor

Abstract

Recommendation systems exist to help users discover content in a large body of items. An ideal recommendation system should mimic the actions of a trusted friend or expert, producing a personalised collection of recommendations that balance between the desired goals of accuracy, diversity, novelty and serendipity. We introduce the Auralist recommendation framework, a system that - in contrast to previous work - attempts to balance and improve all four factors simultaneously. Using a collection of novel algorithms inspired by principles of 'serendipitous discovery', we demonstrate a method of successfully injecting serendipity, novelty and diversity into recommendations whilst limiting the impact on accuracy. We evaluate Auralist quantitatively over a broad set of metrics and, with a user study on music recommendation, show that Auralist's emphasis on serendipity indeed improves user satisfaction.

Auralist: Introducing Serendipity into Music Recommendation

Yuan Cao Zhang[§] Diarmuid Ó Séaghdha[§] Daniele Quercia[§] Tamas Jambor[‡]
[§]Computer Laboratory, University of Cambridge, UK
[‡]Department of Computer Science, University College London, UK
ycz20@cantab.net, do242@cam.ac.uk, dq209@cl.cam.ac.uk, t.jambor@cs.ucl.ac.uk

ABSTRACT

Recommendation systems exist to help users discover content in a large body of items. An ideal recommendation system should mimic the actions of a trusted friend or expert, producing a personalised collection of recommendations that balance between the desired goals of accuracy, diversity, novelty and serendipity. We introduce the *Auralist* recommendation framework, a system that – in contrast to previous work – attempts to balance and improve all four factors simultaneously. Using a collection of novel algorithms inspired by principles of 'serendipitous discovery', we demonstrate a method of successfully injecting serendipity, novelty and diversity into recommendations whilst limiting the impact on accuracy. We evaluate *Auralist* quantitatively over a broad set of metrics and, with a user study on music recommendation, show that *Auralist*'s emphasis on serendipity indeed improves user satisfaction.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Retrieval and Search; H.2.8 [**Database applications**]: Data mining; I.2.6 [**Artificial Intelligence**]: Learning—*Knowledge Acquisition*

General Terms

Algorithms, Experimentation, Human Factors

Keywords

Collaborative filtering, diversification, serendipity, novelty, accuracy, recommender systems, metrics

1. INTRODUCTION

In an era of increasing choice, recommender systems have emerged as an important tool to help consumers manage the dizzying array of options presented by digitised markets and communities. Recommender systems generate prioritised list of unseen items by trying to predict a user's preferences based upon their profile. Such systems can now be seen in numerous applications (recommending, e.g., music, books), and have been shown to aid online sales [21].

Until recently, the vast majority of previous research has focused on improving the accuracy of recommendation: better modelling user preference so as to produce individually more enjoyable items. A growing trend, however, is to consider factors other than accuracy that contribute towards the quality of recommendation [15]. Notions of diversity, novelty and serendipity have been recently explored [8] as objectives that often conflict with the drive for accuracy.

The dangers of an overt focus on accuracy are twofold. Firstly, in failing to consider human desires for variety, discovery and change, accuracy-focused recommenders may forfeit an overall improved user experience, producing boring and ineffective recommendations. Secondly, in considering that recommendations have the power to shape user consumption patterns, there is an aesthetic concern that too much personalisation and pandering to a user's existing tastes harms a user's personal growth and experience. The extreme concept of a "*filter bubble*" describes the idea that users could be trapped in a self-reinforcing cycle of opinion, never being pushed to discover alternative genres or viewpoints [17].

It is important, then, that systems are designed with such alternative qualities in mind. We introduce a series of novel and well-grounded approaches that together compose the *Auralist* recommender, a system that explicitly balances the conflicting goals of accuracy, diversity, novelty and serendipity. Whereas prior research has often focused on these factors individually (as we shall see in our discussion of related work in Section 2), we use a range of metrics to measure all three non-accuracy factors simultaneously. Our methods are introduced as part of a hybrid framework that can combine a variety of algorithms to achieve the desired mix of qualities.

We address the issue of non-accuracy factors by focusing on techniques that simultaneously inject novelty, diversity, and serendipity into the recommendation process. In this way, we hope to actively counteract the constricting effects of personalisation. More specifically, we make three main contributions:

Auralist Framework. Auralist uses hybrid rank-interpolation to combine the output of three constituent algorithms in four different combinations: 1) *Basic Auralist* employs solely a new item-based col-

laborative filtering algorithm called *Artist-based LDA* based on *Latent Dirichlet Allocation* [3] (Section 4.1); 2) *Community-Aware Auralist* combines *Artist-based LDA* with a new algorithm called *Listener Diversity* that promotes artists with "diverse" listenerships (Section 4.2.1); 3) *Bubble-Aware Auralist* combines *Artist-based LDA* with the new *Declustering* algorithm that identifies and counteracts a user's "music bubble" (Section 4.2.2); and 4) *Full Auralist* is the combination of all three algorithms (*Artist-based LDA*, *Listener Diversity*, and *Declustering*) together.

Quantitative evaluation of Auralist. We evaluate the different versions of Auralist using a comprehensive set of metrics that simultaneously assess accuracy, diversity, novelty and serendipity in user recommendation results (Section 5). We find that *Basic Auralist* produces recommendations that are as accurate as those produced by the state-of-the art *Implicit SVD* algorithm [10], and that both *Community-Aware Auralist* and *Bubble-Aware Auralist* greatly improve all three qualities of diversity, novelty and serendipity with differing trade-offs in accuracy.

Qualitative evaluation of Auralist. We conduct a user study to assess the objective qualities of enjoyment, real-world novelty, serendipity and overall user satisfaction (Section 6). We find that *Full Auralist*, although having lower accuracy and individual item enjoyment, is significantly more serendipitous and proves more satisfying overall than *Basic Auralist*.

Before concluding, we discuss the practical implications of our findings that go beyond music recommendation (Section 7).

2. RELATED WORK

Our research builds upon previous work attempting to quantify and measure the additional factors influencing end-user recommendation quality. We also draw inspiration from the implementation of algorithms designed to retain or enhance such qualities.

The techniques we propose belong to a family of model-based techniques for collaborative filtering [22]. Item-based [20] approaches in particular have found use in a number of commercial applications, being pioneered for Amazon.com product recommendations [13]. The possible use of *LDA* for recommendation was touched upon by Blei [3] in his initial formulation of the model, but it was later research by Hoffman [9] that established the use of Latent semantic models (in the form of the PLSI topic model) as tools for collaborative filtering.

Early concepts of novelty and serendipity were described by Herlocker [8] in his seminal survey of recommendation evaluation techniques, with other authors contributing quantitative measures and definitions [29, 27, 24, 16]. More recently, Vargas and Castells [25] attempt to formally unify diversity and novelty in a single evaluation framework. Our framework takes the straightforward approach of selecting a single metric to measure each evaluation independently.

The idea of balancing multiple objectives in recommendation has a strong basis in previous research. Adomavicius and Kwon [1] introduce a re-ranking method of diversity-improvement by applying a number of simple re-ranking algorithms to the output of accuracy-focused collaborative filtering algorithms. The diversity-accuracy tradeoff is controlled by restricting the re-ranking to the *top-N* most accurate items, thus restricting the maximum movement of an item. They show substantial improvements in the number of distinct items recommended in the *top-N* lists of users.

Jambor and Wang [11] frame conflicting objectives as a series of optimisation constraints, for use in a matrix factorisation algorithm. In one such experiment, the authors introduce constraints related to novelty and diversity, which are able to alter the distribution of popular items within a *top-N* list with little impact on accuracy. Ziegler *et al.* [29] use a topic diversification algorithm to rank items according to dissimilarity to preceding items in the recommendation set, integrating this information into recommendations through a rank-interpolation method; an approach that we reproduce, but across a wider set of objectives. A similar strategy appears in Zhou and Kuscik [27], where two graph-spreading algorithms that specialise in terms of accuracy and novelty performance respectively are hybridised in order to produce a recommender with the properties of both. Their results suggest that interpolation is indeed a viable method of producing balanced recommendations.

It is unclear, however, if multiple properties in addition to accuracy can be promoted at once. We present next a detailed description of the three properties of diversity, novelty and serendipity (Section 3) and a selection of algorithms that attempt to incorporate those properties in the recommendation process (Section 4).

3. WHY ACCURACY IS NOT ENOUGH

Before defining the three properties, we introduce two widely-used measures of accuracy and explain when they fall short of what is needed to measure the effectiveness of recommender systems.

Traditionally, recommendation quality is measured using one of a number of accuracy metrics, which assess how well the predictive output of a system matches a proportion of known, withheld items for each user. Examples of accuracy metrics include average *Top-N Recall* and the average *Rank* score proposed by Hu *et al.* [10]. Recall-based metrics measure the proportion of a user's test-set that appear within a *Top-N* recommendation list – a *Top-20* list that successfully includes 5 out of 10 test-set items will score a single-user recall of 0.5. These metrics we describe using a set of common symbols in Table 1.

$$\overline{Top\text{-}20\ Recall} = \frac{1}{|S|}\sum_{u\in S}\frac{|R_{u,20}\cap W_u|}{|W_u|} \tag{1}$$

The average *Rank* score measures the average percentage rank of withheld items in a user's history, weighted by (positive) preference (for boolean preferences, the average *Rank* is simply the average rank of all history items). Unlike $\overline{Top\text{-}20\ Recall}$, $\overline{Rank}$ assesses how accurate a system is at modelling a user's entire history, not just the most easily recommended items. It also takes into account the full recommendation ranking, which complements recall-based assessments of only the top (observable) items:

$$\overline{Rank} = \frac{1}{|S|}\sum_{u\in S}\frac{\sum_{j\in N}P_{j,u}rank_{j,u}}{\sum_{j\in N}P_{j,u}} \tag{2}$$

S	Set of all users
P	Preference matrix, where $P_{i,u}$ is the preference given by user u for item i (with our boolean preference system, $P_{i,u} = 1$ if the item is preferred by the user, and 0 otherwise)
R	*Top-N* function, where $R_{u,n}$ gives the top n recommended items for user u
N	Set of all items
W_u	Withheld item history of user u (set not used for training)
H_u	Non-withheld item history of user u
pop_i	Fraction of (all) preferences directed at item i
T	Set of *LDA* topics (the number of topics is 200).
L	*LDA* item-topic matrix, where $L_{i,t}$ represents the composition proportion assigned to topic t for item i
C_i	Number of artist i's unique listeners

Table 1: Symbols.

$$rank_{j,u} = \frac{\text{Index of item } j \text{ in ordered list for } u}{|N|}$$

Other metrics, such as *RMSE* (which measures the standard error of predicted preferences) have also seen widespread use. The above metrics, however, are better suited to handle the boolean nature of our dataset, which is described further in Section 5.1. The use of accuracy as a performance metric is well-grounded; previous user studies, such as the one conducted by Swearingen and Sinha [23] indicate that recommendation accuracy is a primary factor behind user satisfaction with recommendation systems. This has led to a focus on improving accuracy in recommendation algorithms; state-of-the-art systems score very highly indeed [10]. However, this is not to say that accuracy alone guarantees satisfactory recommendations.

There is a growing argument that factors other than accuracy also influence recommendation quality [15, 8, 29, 5, 19]. Recommendation techniques that focus purely on accuracy may neglect such alternative qualities and produce recommendations that appear superficially "good" but are in fact inferior in terms of actual user satisfaction. An extreme example of this may be a recommendation set consisting of entirely Beatles songs - the recommendations themselves may be accurate, but users will rapidly become bored with a collection of such similar and generic items.

To fix this problem, the three criteria of diversity, novelty and serendipity have been introduced by researchers. These assess the major factors influencing recommendation satisfaction alongside accuracy. Unlike previous work which has often considered only one or two such factors, we measure our improvements against a comprehensive assessment of each. To better characterise likely usage scenarios, all our metrics are applied to the *Top-20* recommendation list for each user, representing the fact that in a realistic application users are unlikely to be exposed to anything below the very top of the rankings.

Diversity represents the variety present in a list of recommendations. A diverse list can be seen to counteract user satiety with (homogeneous) recommendations. The aforementioned all-Beatles recommendation list, for example, is much less diverse than a list containing a wider assortment of artists. Previous research has shown that users will actively choose less-preferred items in an effort to improve the variety of consumption [19, 29]. We measure diversity through the *Intra-List Similarity* metric introduced by Ziegler *et al.* [28, 29], using (binary) cosine similarity (*CosSim*) to judge the similarity between items. *Intra-List Similarity* essentially sums the pairwise similarity of all items in a set (simplified in our case due to a symmetric similarity measure). Intuitively, the greater the proportion of preferring users two items have in common, the greater the similarity value *CosSim* will register. A recommendation list with groups of very similar items will score a high intra-list similarity compared to a list that has more dispersed and diverse recommendations.

$$\overline{Intra\text{-}List\ Similarity} = \frac{1}{|S|}\sum_{u\in S}\sum_{i,j\in \mathrm{R}_{u,20}, j<i} CosSim(i,j) \quad (3)$$

$$CosSim(i,j) = \frac{\#\text{ users who like both } i \text{ and } j}{\sqrt{\#\text{ prefs in } i} \times \sqrt{\#\text{ prefs in } j}} \quad (4)$$

Novelty can be seen as the ability of a recommender to introduce users to items that they have not previously experienced before in real life (such experiences may be outside the system itself; e.g., music listened to whilst not on a computer). A recommendation that is accurate but not novel will include items that the user enjoys, but already knows of. A limited proportion of such recommendations has been shown [23, 18] to have a positive, trust-building impact on user satisfaction, but it can also be seen that to be useful a recommender needs to suggest previously unknown items. We measure novelty with a metric previously introduced by Zhuo and Kuscik [27]:

$$\overline{Novelty} = \frac{1}{|S|}\sum_{u\in S}\sum_{i\in \mathrm{R}_{u,20}} \frac{\log_2 pop_i}{20} \quad (5)$$

This novelty metric quantifies the average information content of recommendation events – higher values mean that more globally "unexplored" items are being recommended. Under the assumption that the likelihood a user has experienced an item is proportional to its global popularity, this serves an approximation of true novelty. We measure actual novelty on an individual basis in our user study (Section 6).

Serendipity represents the "unusualness" or "surprise" of recommendations. Unlike novelty, serendipity encompasses the semantic content of items, and can be imagined as the distance between recommended items and their expected contents. A recommendation of *John Lennon* to listeners of *The Beatles* may well be accurate and novel, but hardly constitutes an original or surprising recommendation. A serendipitous system will challenge users to expand their tastes and hopefully provide more interesting recommendations, qualities that can help improve recommendation satisfaction [23]. We assess serendipity through a new *Unserendipity* metric, which uses *CosSim* to measure the average similarity between items in a user's history H_u and new recommendations. Lower values indicate that recommendations deviate from a user's traditional behaviour, and hence

are more surprising:

$$\overline{Unserendipity} = \sum_{u \in S} \frac{1}{|S||H_u|} \sum_{h \in H_u} \sum_{i \in R_{u,20}} \frac{CosSim(i,h)}{20} \quad (6)$$

This metric bears some similarities to the distance-based novelty family of metrics seen in Vargas *et al.* [25].

4. THE AURALIST FRAMEWORK

At heart, the *Auralist* framework is an experiment in combining distinctive recommendation algorithms to improve overall (serendipitous) performance. Here, we introduce three techniques for generating recommendation rankings (*Artist-based LDA*, *Listener Diversity* and *Declustering*) that are paired to create different flavours of *Auralist*. *Basic Auralist* (Section 4.1) incorporates the so-called *Artist-based LDA* technique and is intended as a standalone recommender system. *Community-Aware* (Section 4.2.1) and *Bubble-Aware* (Section 4.2.2) *Auralist* versions interpolate *Artist-based LDA* with *Listener Diversity* and *Declustering* rankings respectively. These versions combine small elements of a serendipity-focused algorithm to reorder the basic algorithm's recommendations. A final *Full Auralist* recommender combines all three sub-algorithms.

4.1 Basic Auralist

Basic Auralist is an item-based recommender system that employs *Latent Dirichlet Allocation* as a technique for computing item features. We call this approach *Artist-based LDA* and present it next.

LDA has been used traditionally in topic-modeling, being a fully generative model for document production [3]. Under this framework, words within a large document set can be clustered into topics based upon co-occurrence, each topic being a probabilistic distribution over word tokens. A "topic composition vector" can then be determined for each document, indicating the estimated level of influence each "topic" would have if the document were to be generated using the *LDA* model. Both topic clustering and document composition can be computed stochastically using the *Gibbs Sampling* algorithm (described in Griffiths and Steyvers [7]) in an unsupervised manner over a training dataset.

Our approach applies *Gibbs Sampling* to the unary preferences of our *Last.fm* dataset (described in Section 5.1) using the MALLET toolkit [14]. We note two straightforward means of framing user-artist preferences in a manner suitable for *LDA*. In a *User-based LDA* model, users are treated as *LDA* documents and preferred artists as words. This produces a series of artist topics corresponding roughly to artist genre, but does not tell us anything about the artists themselves. Conversely, the inverse *Artist-based LDA* model treats artists as documents and preferring users as words, producing a fixed-length topic composition vector for each item. Topics in the artist-based model can be imagined to represent *user-communities*, clustering together users with similar preferences. Topic vectors thus represent the distribution of the listener base of an artist, and can be used to characterise them. We then define a *LDA* similarity metric as the (real-valued) cosine similarity between **artist** topic vectors:

$$LDASim(i,j) = \frac{\sum_{t \in T} L_{i,t} \times L_{j,t}}{\sqrt{\sum_{t \in T} (L_{i,t})^2} \sqrt{\sum_{t \in T} (L_{j,t})^2}} \quad (7)$$

This similarity metric is then used directly for item-based recommendation by defining $Basic(u,i)$, which is the score that user u associates to item i:

$$Basic(u,i) = \sum_{h \in H_u} LDASim(i,h) \quad (8)$$

All artists can be sorted (in descending order) by the sum-total of their similarity with items in a user's existing history [13]. This produces a $rank_{Basic,u,i}$ for each item i, with the most similar items awarded the smallest percentage ranks.

By generalising user "topics", we aim to both smooth the data and generate less obvious and more serendipitous recommendations compared to more naïve techniques, as connections can now be made through similar, but not directly related users. We note also that *Basic Auralist* built on *Artist-based LDA* inherits two main benefits common to *model-based* recommenders: faster online performance (if item-item similarities are not precomputed) and a compact semantic representation of the data (in this case, in terms of listener composition). This semantic representation is exploited in the following subsection to further influence recommendation.

4.2 Two hybrid versions of Auralist

To increase the novelty, diversity, and serendipity of *Basic Auralist*'s recommendations, we combine *Artist-based LDA* recommendation with two new algorithms. The first is called *Listener Diversity* and aims to prioritise for recommendation artists with particularly diverse listener communities, encouraging users to explore beyond a given niche. The combination of *Artist-based LDA* and *Listener Diversity* is called *Community-Aware Auralist* (Section 4.2.1).

The second algorithm is called *Declustering* and aims to determine a user's "musical bubbles" (clusters of artists that the user listens to) and th[illegible]commend artists outside established cluster groups (he[illegible] *Declustering*). The combination of *Artist-based LDA* and *Declustering* is called *Bubble-Aware Auralist* (Section 4.2.2).

We combine the different *Auralist* algorithms by merging their individual rank outputs. One way of doing so is to produce a hybrid score for each item (artist) [29]. Intuitively, the hybrid ranking score of an item i can be taken as a linear interpolation of the percentage [0,1) rank the item has in the output of each of the contributing algorithms. A set of interpolation coefficients λ_a over a set of algorithms A controls the influence of each individual algorithm. In the case of the generalised *Full Auralist* recommender, we have three λ coefficients governing an algorithm set A that includes *Artist-based LDA*, *Listener Diversity* and *Declustering*.

$$Hybrid(u,i) = \sum_{a \in A} \lambda_a (rank_{a,u,i})$$

The final recommendation output consists of the item list sorted by the hybrid rank score. The "hybridisation" of recommendation allows an accuracy-focused *Basic Auralist* to be combined with small proportions of diversity or serendip-

ity promoting algorithms, in order to improve the overall balance of qualities.

4.2.1 Community-Aware Auralist

Community-Aware Auralist introduces the *Listener Diversity* metric for artists, which is used to produce a ranked list of the most diverse artists. This list is blended with *Basic Auralist* to promote more diverse artists in recommendation.

We recall that for the *Artist LDA* model, topics are formed over "user communities", groups of users that share common item preferences. An artist in the *LDA* recommender is represented by a vector of topic proportions indicating how listeners of that artist are distributed amongst these "user communities".

Such a representation offers us a unique perspective on the demographics of listeners, not visible when observing the raw vector of preferences. Certain artists, whilst popular in their own right, might have a listener base concentrated in only a few user communities, whereas the listeners of another artist might be more widely distributed.

Given that a *LDA* topic vector is a probability distribution summing to 1, we use the entropy of such a distribution to measure its skewedness. A distribution focused on only a few outcomes will score a less negative entropy; a more evenly and widely distributed event will produce a greater negative entropy. We thus introduce the *Listener Diversity* of an artist i as the entropy over its topic distribution:

$$\textit{Listener Diversity}(i) = -\sum_{t \in T} L_{i,t} \log_2(L_{i,t})$$

What does Listener Diversity represent in terms of recommendation quality? Intuitively, we can imagine it as being a measure of nicheness - how polarising a given performer is. A Ukrainian bagpipe metal band[1] is unlikely to spark broad appeal compared to, say, *The Beatles*. However, that is not to say that the former should not be recommended, if the user belongs to the limited demographic following that style of music. In the context of serendipitous recommendations, however, we seek to expand a user's music taste beyond that of his comfort zone. A strategy for this would be to highlight more diverse artists that include a user's established music communities, but also introduce elements of ones the user may be unfamiliar with. This balance can be achieved by interpolating the output of a *Listener Diversity*-sorted list with that of a conventional accuracy-focused algorithm; the former boosting the rank of more diverse artists whilst the latter ensures that ranking artists are still enjoyable:

$$\textit{Community}(u,i) = (1-\lambda) rank_{Basic,u,i} + \lambda rank_{Diversity,i}$$

Analysis of *Listener Diversity*'s relationship with other factors shows that *Listener Diversity* tends to bias towards globally popular artists. This should be unsurprising, as such artists will garner more exposure and attract a naturally wider fan base. We compensate for this by discounting an artist's original *Listener Diversity* with a popularity-diversity regression function, highlighting artists that are diverse for their popularity level (popularity being the number C_i of the artist's unique listeners). The resulting adjusted *Listener Diversity* is:

$$\textit{Listener Diversity}'(i) = \textit{Listener Diversity}(i) - \textit{Offset}_{pop}(i) \quad (9)$$

In our dataset (which will be described in Section 5.1), following a linear regression, we find the following coefficients for $\textit{Offset}_{pop}(i)$:

$$\textit{Offset}_{pop}(i) = 0.462 \log(C_i) - 1.326. \quad (10)$$

4.2.2 Bubble-Aware Auralist

As a counterpart to *Listener Diversity*, we introduce a graph-based algorithm termed *Declustering*. *Declustering* produces a ranked list of the least "clustered" or "boring" items for a user and is interpolated with *Basic Auralist* to form *Bubble-Aware Auralist*:

$$\textit{Bubble}(u,i) = (1-\lambda) rank_{Basic,u,i} + \lambda rank_{Declustering,u,i}$$

We compute *Declustering* scores over what we call the "Artist Graph". Formally, this is a graph $G = (N, E)$ where each node $i \in N$ is an artist and edges $(i, j, weight) \in E$ are drawn between artists that have non-zero similarity, according to a similarity metric $weight = sim(i, j)$ computed with the previously defined *LDASim* (Equation 7). Intuitively, framing recommendations in this format allows us to apply network-based analysis techniques to (prospective) items. Such a model has been used previously by Celma *et al.* [5, 4] to investigate the long-tail properties of music recommendation in terms of network links. We further introduce the idea of a user's "*local preference graph*", which can also be seen as a user's "music bubble". This is the subgraph $G_u = (H_u, E_u)$ of the artist G consisting only of the nodes $i \in H_u$ that are found in the preference history of the user.

The *Declustering* algorithm attempts to identify nodes that lie on the edge of clusters in a user's graph, avoiding heavy concentrations of previous activity ("boring" recommendations) whilst still maintaining overall similarity. In this way we hope to help users expand their music taste, literally pushing the boundaries of the region their behaviour occupies in the feature-space. This recommendation strategy is motivated by concepts of social network theory such as clustering and brokerage [6].

In analysing the Artist Graph, we find that the *Last.fm* dataset has a power-law distribution of node degrees, suggesting it may have "small-world" properties [26]. This implies a graph structure similar to that of a social network, with nodes being clustered around a series of high-degree "hubs". Therefore, we employ a metric commonly used in social-network analysis to measure how clustered nodes in a network are. The clustering coefficient of a node i is defined as:

$$\textit{Clustering}(i) = \frac{2 \times |\{(j,k) \in E_u | j,k \in \text{neighbours}(i)\}|}{|\text{neighbours}(i)| \times (|\text{neighbours}(i)| - 1)} \quad (11)$$

where *neighbours(i)* is the set of nodes that are neighbours of item i in the local preference graph. The clustering coefficient of a node measures the proportion of possible interconnections that exist amongst neighbours of a node. A node with a high clustering coefficient is surrounded by tightly interconnected nodes (i.e., in the centre of a near-clique) whereas a node with a lower clustering coefficient might have neighbours split between multiple clusters or lie on the edge of an existing cluster.

[1] `www.holyblood.com.ua`

```
foreach item pair j, k in H_u do
    total+= LDASim(j, k);
    count+=1;
end
avgSim = total/count;
foreach item i in candidate set for user u do
    foreach item pair j, k in neighbours(i) do
        if LDASim(j, k) > avgSim then
            edgeTotal+=1;
        end
        edgeCount+=1;
    end
    cluster(i) = edgeTotal/edgeCount;
    recommendations.add(cluster(i), i);
end
recommendations.sortBy(cluster, ascending);
```

Algorithm 1: Pseudo-code for *Declustering*.

The *Declustering* algorithm (Algorithm 1) considers in turn all the prospective recommendations for a user. For each item, it temporarily adds the item to the graph and computes the clustering coefficient of that node with respect to the existing elements of the user's local preference graph. The output of the algorithm is an ordered list of the most "cluster-avoiding" artists in the candidate set, which (as with *Listener Diversity*) can be interpolated with a conventional recommender to apply counter-clustering pressure to recommendation items. As the clustering coefficient operates over unweighted edges, we threshold edges in the local preference graph according to whether a similarity weight indicates unusual significance. An *LDASim* weight that exceeds the average for a particular user is considered significant for the purposes of computing the clustering coefficient. This has the effect of removing the majority of (weak) edges in the graph. We can also adjust the threshold in terms of standard deviations from the average; this affects the sensitivity of cluster detection and should be experimentally determined. *Declustering*'s emphasis on low clustering scores reflects the desire to dissuade the recommendation of artists that are too deeply embedded in a genre cluster with respect to a particular user. Such artists are likely accurate but not serendipitous, being too similar to a great many of user's existing artists.

In the next section, we prove the effectiveness of our techniques by apply to them the evaluation metrics introduced in Section 3.

5. EVALUATION

The goals of our evaluation are to assess: 1) to which extent the *Auralist* framework produces diverse, novel, and serendipitous recommendations; 2) at which cost for accuracy *Auralist* produces such recommendations; and 3) the best combination of algorithms that produces an overall more satisfying recommender. To meet these goals, we employ the suite of metrics described in Section 3 to quantitatively measure the performance of both a set of baseline recommenders as well as various interpolations of our serendipity-enhancing techniques.

5.1 Dataset

Our experiments are conducted over a 360k *Last.fm* user[2] dataset [12], collected by Òscar Celma in 2008 [3]. This contains the `user.getTopArtists()` output for each user from the *Last.fm* API, which is a list of previously listened-to artists and play-counts derived from both *Last.fm*'s online radio services and media player plugins.

In contrast to other publicly available datasets, the *Last.fm* dataset consists of implicit observations of user preference through prior behaviour. This means that there is no explicit ratings scale associated with preferences and that preferences themselves can be considered noisy - track metadata may be incorrect, songs may be left on loop/shuffle and user history lengths will vary. We clean the dataset to remove non-artist items, misspelled artist names and extremely unpopular artists, leaving us with 48,988 possible recommendation items. We take our implicit preferences to be **unary** (1 or nothing), which lends itself well to the processing techniques we introduced in Section 4.

5.2 Basic Auralist Recommendation

We evaluate the effectiveness of *Artist-LDA* recommendation method against the state-of-the art *Implicit SVD* method introduced by Hu, Koren and Valinsky [10]. We adapt this model to incorporate the implicit artist playcount as a confidence weight in the matrix factorisation cost function. Metrics are computed over random subsamples of 35k users; larger samples only marginally improve performance. 20% of each user's preferences were randomly withheld as a training sample. One feature of *LDA* that we leverage is the fact that Gibbs Sampling runs relatively quickly even on large user samples, compared to other model-based techniques. We thus bootstrap the *LDA* topic training step with the full 360k user dataset, which completes 1000 iterations in under an hour on an Intel Core^{TM}i5 2.8GHz processor.

Our experimental results are reported in Table 2 and show that *Basic Auralist* produces the most overall accurate rankings for user histories ($\overline{Rank} = 0.0194$) whilst *Implicit SVD* produces the highest $\overline{Top\text{-}20\ Recall}$ scores (0.174). Both algorithms score comparatively in diversity ($\overline{Intra\text{-}List\ Similarity}$), whereas *Implicit SVD* has improved serendipity and *Basic Auralist* has slightly improved novelty.

The combination of accuracy scores seem to indicate that *Implicit SVD* does a better job of including items in the *Top-20* list. However, it may be argued that for the use-case of *serendipitous* recommendation, a high recall is not necessary; recall indicates that similar, already known items are being placed in the *Top-20* list, displacing the recommendation of novel items. By contrast, *Basic Auralist* broadly characterises what a user has previously liked ($\overline{Rank}$) without being overtly sycophantic. Interestingly, of the items *Implicit SVD* does recommend, the registered $\overline{Unserendipity}$ is somewhat lower, implying that the generalisation of matrix factorisation does result in some less obvious recommendations as well. We exceed this serendipity value with later versions of *Auralist*.

5.3 Hybrid versions of Auralist

[2] `http://www.last.fm`

[3] `http://www.dtic.upf.edu/~ocelma/MusicRecommendationDataset/`

	$\overline{Rank}$	$\overline{Top\text{-}20\ Recall}$	$\overline{Intra\text{-}List\ Similarity}$	$\overline{Novelty}$	$\overline{Unserendipity}$
Basic Auralist	**0.019** ±0.0004	0.157 ±0.004	**14.4** ±0.2	**11.8** ±0.06	0.060 ±0.0004
Implicit SVD	0.039 ± 0.0008	**0.174** ±0.002	14.7 ±0.1	10.9 ±0.03	**0.046** ±0.0002
Community-aware(λ=0.05)	0.023 ±0.02	0.030 ±0.0009	3.4 ±0.06	17.2 ±0.1	0.047 ±0.0003
Bubble-aware(λ=0.2)	0.021 ±0.0002	0.029 ±0.0006	3.4 ±0.05	14.2 ±0.1	0.035 ±0.0002
Full Auralist	0.025	0.008	1.54	17.3	0.039

Table 2: Performance results for *Basic Auralist*, the state-of-the-art *Implicit SVD*, and *Full Auralist*.

Figures 1 and 2 show the performance results for *Community-Aware* and *Bubble-Aware Auralist*. Table 2 also includes their performance at points of interest along the λ curve (note at λ=0, both algorithms reduce to *Basic Auralist*).

Given that both hybrid versions of Auralist attempt to bias towards serendipitous recommendations at the expense of more "easily accurate" items, it should be unsurprising that both exhibit an accuracy-serendipity trade-off. More interestingly, both methods increase novelty and diversity, and do so at different rates.

As the *Listener Diversity* interpolation increases, *Community-Aware Auralist*'s rapid improvements in non-accuracy scores (Figures 2(a), 2(b), 2(c)) are tracked by decays in recall (Figure 1(b)) and to a lesser extent $\overline{Rank}$ (Figure 1(a)), tailing off at higher proportions. *Community-Aware Auralist* hence represents a direct trade-off between accuracy and non-accuracy performance, with the most activity occurring in the $0 < \lambda < 0.05$ range of Figures 1 and 2. Compared with the other graphs, *Community-Aware Auralist* maintains a consistently sizable lead over *Bubble-Aware Auralist* in terms of novelty (Figure 2(b)), likely due to the popularity correction $Offset_{pop}$ we introduced in Section 4.2.1 (Equation 10).

As with *Community-Aware Auralist*, the *Bubble-Aware Auralist*'s performance curves for serendipity, novelty and diversity track that of $\overline{Top\text{-}20\ Recall}$. Unlike *Community-Aware Auralist*, however, *Bubble-Aware Auralist*'s $\overline{Rank}$ decays at a much slower rate, and the performance curves possess sigmoid-like qualities, experiencing the greatest rate of change after about $\lambda = 0.1$ and diminishing returns afterwards. We propose that this is the point when the *Declustering* algorithm is able to successfully overcome the bias towards recommendations embedded within preference clusters and is able to successfully recommend cluster-bordering items. The nature of the $\overline{Rank}$ curve indicates that the bulk of this benefit can be achieved without an overwhelming effect on overall accuracy, suggesting that *Bubble-Aware Auralist* may be able to supply "almost free" serendipity, diversity and novelty.

Bubble-Aware Auralist manages to surpass *Community-Aware Auralist* in terms of serendipity relatively quickly ($\lambda \sim 0.15$), continuing to improve even after *Community-Aware Auralist*'s performance begins to plateau. This suggests that the serendipity improvement is not merely incidental (i.e., from declining accuracy), and is actively being promoted by *Declustering*.

To sum up, these findings indicate that *Community-Aware Auralist* is best used at smaller interpolations (0–0.05) as a roughly even trade between accuracy and non-accuracy qualities and as a broad stroke in changing the (accuracy/non-accuracy) focus of a recommender. They also suggest that using *Bubble-Aware Auralist* during the peak rate of change before significant $\overline{Rank}$ penalties (Figure 1(a)) accrue can improve non-accuracy qualities at very little cost. At $\lambda = 0.2$, a mere 0.7% increase in average history rank is accompanied by a 77% decrease in $\overline{Intra\text{-}List\ Similarity}$, 20% increase in novelty and a 42% decrease in measured unserendipity. Overall, both methods prove to be able to successfully improve diversity, novelty and serendipity.

6. USER STUDY

Alongside our quantitative measurements, we further conduct a user study to validate the effectiveness of *Auralist* (and indeed, serendipity-orientated recommendation in general) in real-life situations. We measure the perceived serendipity, enjoyment, novelty and overall qualitative satisfaction associated with a refined version of the hybrid recommender.

The *Full Auralist* recommender combines *Artist-LDA*, *Listener Diversity*, and *Declustering* in proportions of $\lambda_{1,2} = (0.03, 0.20)$ respectively (motivated by the results of the previous section) and demonstrates overall superior non-accuracy performance compared to our previous methods (Table 2):

$$\begin{aligned} Full_{u,i} =& (1 - \lambda_1 - \lambda_2) rank_{Basic,u,i} \\ &+ \lambda_1 rank_{Diversity,u,i} + \lambda_2 rank_{Declustering,u,i} \end{aligned}$$

6.1 Experimental Method

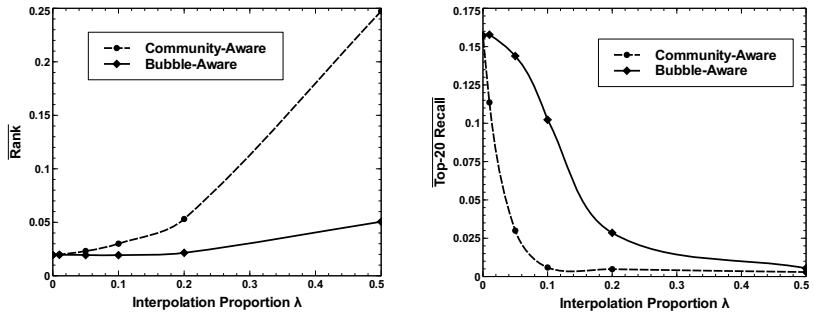
The user study involved 21 participants, the majority of which are current university students. This included a mix of under/post graduates and men/women between the ages of 18-27, of varying nationalities. Each participant was asked to name six pre-2008 artists that represented his/her music tastes, which were used as "seed" histories for recommendation. Volunteers suggested a very wide range of artists, across many musical genres. Users were then presented with two (unlabelled) *Top-20* recommendation lists, generated by *Basic Auralist* and *Full Auralist* respectively. The lists were presented in a randomly determined order for each participant.

Users were instructed to listen to at least two 30-second song samples from each unknown artist and to fill in an accompanying survey[4]. Survey questions assess individually for each recommendation how enjoyable (Dislike the song... Will definitely listen again), serendipitous (Exactly what I listen to normally... Something I would never have listened to otherwise) and novel an artist is. The former two are assessed using 5-point Likert scales, whereas novelty is multiple choice.

6.2 User Ratings

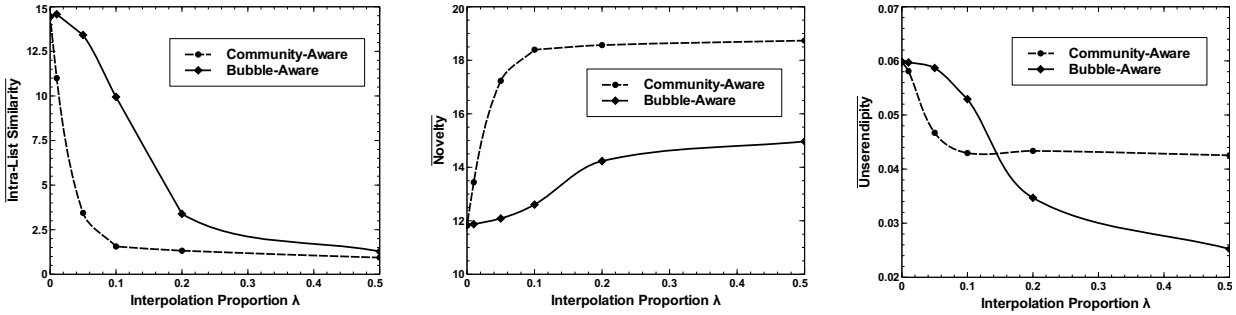
For each user, we compute the average enjoyment and serendipity ratings given to the artists in each list. The results are summarised in Table 3 and show that there is a substantial difference in scores given to *Basic Auralist* and

[4] `http://tinyurl.com/ycz20`



(a) $\overline{Rank}$ accuracy (b) $\overline{Top\text{-}20\ Recall}$ accuracy

Figure 1: Accuracy performance of *Community-Aware Auralist* and *Bubble-Aware Auralist* as the contribution λ of the corresponding serendipity-enhancing technique increases.



(a) Diversity (b) Novelty (c) Serendipity

Figure 2: Diversity, novelty, and serendipity performance of *Community-Aware Auralist* and *Bubble-Aware Auralist* as the contribution λ of the corresponding serendipity-enhancing technique increases.

	Basic Auralist	**Full Auralist**
$\overline{\text{Serendipity Rating}}$	2.08 ($\pm$0.77)	2.96 ($\pm$0.69)
$\overline{\text{Enjoyment Rating}}$	4.21 ($\pm$0.54)	3.82 ($\pm$0.53)
$\overline{\text{\# Useful Recommendations}}$	2.90 ($\pm$2.61)	5.86 ($\pm$3.05)
$\overline{\text{\# Serendipitous Recomm.s}}$	1.81 ($\pm$1.86)	4.14 ($\pm$2.90)
$\overline{\text{\# Familiar Recomm.s}}$	12.62 ($\pm$4.14)	7.43 ($\pm$5.07)

Table 3: Summary of the results from our user study. Values in brackets are standard deviations.

Full Auralist. Compared to the basic accuracy-focused system, *Full Auralist* manages to score much higher in terms of *Serendipity* (+0.88), but sacrifices a proportion of average *Enjoyment* ($-$0.39) in doing so. A plot of the difference between these two variables can be seen in Figure 3(a).

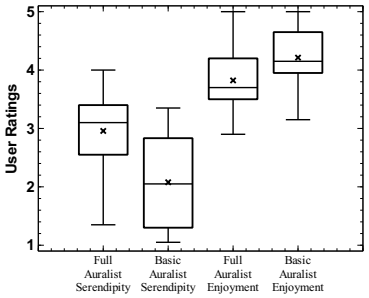
Whilst the variance of reported results may appear high, we recall that the study was conducted as a repeated measures experiment. Therefore, we test the significance of a findings using a one-tailed pairwise *t-test*. Our tests show that *Full Auralist* does indeed exhibit greater serendipity ($p = 0.00002$) and reduced accuracy ($p = 0.004$) compared to *Basic Auralist*.

In addition to raw serendipity/enjoyment scores, we also measure the average number of *Useful*, *Serendipitous* and *Familiar* recommendations issued to each user by the recommenders. We classify as *Useful* any recommendation that is not "Already Known" to the user, and is rated a 4 or 5 in enjoyment ("Already Known" recommendations with 4/5 in enjoyment are instead classified as *Familiar* recommendations). *Serendipitous* recommendations are those *Useful* recommendations that satisfy the additional requirement of being rated a 4 or 5 in serendipity. Useful items represent successful recommendations made to the user, whilst Serendipitous items detail how many managed to both satisfy the user and expand his/her tastes at the same time. Familiar items represent the "trust-building" items described in Swearingen [23] that do not increase utility but improve user satisfaction with the system.

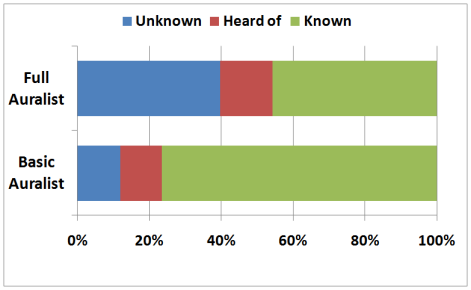
Full Auralist is shown to improve significantly on the basic version in terms of the number of Useful and Serendipitous recommendations, with *t-test* confidence values of $p < 0.001$. Indeed, *Full Auralist* produces on average double the number of Useful and Serendipitous recommendations compared to *Basic Auralist*, with the proportion of Serendipitous artists within the set of Useful artist being higher as well (71% compared to 62%). Despite this, *Full Auralist* still manages to produce a substantial number of Familiar recommendations (7.43). The overall novelty of *Full Auralist*'s recommendations is vastly superior - Figure 3(b) shows that the system reduces the number of "Already Known" artists in recommendation lists by an absolute percentage of over 25%.

6.3 User Satisfaction

Quantitative analysis of survey results have shown that, whilst featuring a reduction in perceived *Enjoyment*, *Full*



(a) *Serendipity* and *Enjoyment* ratings



(b) Fraction of novel recommendations

Figure 3: Results for user satisfaction. (a) *Serendipity* and *Enjoyment* user ratings for *Basic Auralist* and *Full Auralist* on a 1-5 Likert scale. (b) Fraction of recommended artists that are *"Previously Known"*, *"Heard of, but not listened to"* and *"Completely Unknown"* for *Basic Auralist* and *Full Auralist*.

Auralist significantly improves upon perceived *Serendipity* and *Novelty*, as well as the overall number of *Useful* recommendations. We show below a selection of comments participants made with regards to which recommendation algorithm they thought was more satisfying:

> *"[Full Auralist] was more satisfying because it introduced me to new artists. [Basic] was filled entirely with new artists which, while very good, were things that I listened to all the time on a regular basis. [Full Auralist] had artists that were of the same quality of those I listen to but which I'd never heard of."*

> *"I found [the Full Auralist list] more surprising than [Basic]. Most artists I had not heard of (which is what I prefer). Listening to them gave me at least five new artists I could look into and use in the future."*

> *"While I enjoyed the songs on the [Full Auralist] list less, I liked that there was more new music on it than the first list. So I'm going to say that I preferred the [Full Auralist] list."*

> *"[The Basic list was better], more familiar music & more my taste, although [Full Auralist] introduced me to a few good bands."*

> *"[The Full Auralist list] was way too jazzy, and had very few artists I connected with immediately. While [the Basic list] had a vast majority of artists I knew well and have opinions of, the few unknowns were really very congenial."*

In analysing these and other comments, we see two broad consensuses amongst the opinions of participants. A majority(12) of users prefer *Full Auralist*, appreciating the novelty and serendipity of the recommendations made. A substantial minority(7), however, prefer the baseline system due to its comparatively better modelling of their own tastes. Two gave neutral preferences. Whether a user falls into the first or second camp may well depend on three main factors - the user's prior convictions, emotional state, and social context, as recent work has suggested [2].

The dichotomy in preference seems to suggest that an adaptive recommendation system, where users can individually tune the level of "wildness" in recommendations, may find success. This could be offered as a series of recommendation lists, or perhaps be controlled by a sliding scale. In both cases, the hybrid model of recommendation would be particularly effective, as different levels of serendipity can be implemented simply by adjusting the interpolation parameters.

7. DISCUSSION

Our user study shows that *Full Auralist* produces significantly more serendipitous recommendations at the cost of some accuracy. We also show that, despite the reduced accuracy, a large number of participants expressed greater satisfaction with *Full Auralist*'s recommendations. These conclusions are consistent with previous user studies [29, 23, 19]. In particular, we support Ziegler [29] and Ratner [19]'s findings that users are willing to sacrifice some amount of accuracy for improved novelty/diversity/serendipity performance, and that such systems are more satisfying overall. Qualitative comments seem to indicate that serendipity is usually, but not consistently, a positive contributor to this.

The nature of this experiment also demonstrates that Auralist functions well as a practical recommender, even with the "cold-start problem" of limited initial history. It is likely that much of the results variance comes from users' choice of initial artists, with some users suggesting a wider range of genres. Additional history data, perhaps pulled from a *Last.fm* profile, would allow us to better model a user's preferences and thus generate both more accurate and more serendipitous items.

Perhaps our most interesting discovery is that novelty, diversity and serendipity can be improved simultaneously, without any apparent trade-off between the three qualities. One may argue that this is because all three benefit from a departure from pure accuracy – all three qualities, though different, represent facets of a notion of "discovery" that diametrically oppose the notion of "familiarity" that accuracy represents. This does not mean individual qualities cannot be emphasised, however - hybridising a popularity-sorted ranking list will primarily improve novelty, whereas a topic diversification method (such as Ziegler's [29]) will mostly improve diversity.

Our algorithms represents a direct attempt at countering what appears to be an increasing trend by websites and social media to filter what people see - recommending only "safe" items by clustering like-minded users [17]. This behaviour is concerning because it prevents established ideas and norms from being challenged, fostering partisanship and impeding mutual understanding. We hope that even as algorithms are becoming more accurate, additional consideration is given to ensuring this accuracy is also used to introduce users to new content. Despite being designed for music recommendation, *Auralist* has the potential to be adapted for a great many other fields; practically, the *Declustering* algorithm could be readily applied to many existing item-based recommendation algorithms (with domain-dependent similarity metrics). By mapping out a user's preference space and deliberately trying to expand it, one could effectively promote personalised exploration balanced with satisfaction.

8. CONCLUSION

We introduced *Auralist* as a novel recommendation framework that generates diverse, novel and serendipitous recommendations, at a slight cost to accuracy. To aid quantitative analysis, we described a series of metrics designed to assess both accuracy and the three additional qualities of diversity, novelty and serendipity.

We further presented two novel serendipity-enhancing techniques that can be combined with existing ranking methods through "hybridisation". Both *Community-Aware Auralist* and the *Bubble-Aware Auralist* prove to effectively boost novelty, diversity and serendipity scores, with the latter offering a better trade-off with regards to accuracy.

Through a user study on the *Full Auralist* recommender employing all three techniques, we conclusively show that our methods are able to produce more serendipitous recommendations. In addition, despite a noted decrease in the average enjoyment of artists, we show that our serendipity-enhancing techniques improve overall user satisfaction.

We believe our findings are valuable for any kind of consumer-facing recommendation system, where a user's previous history may increasingly constrain their recommendations. Our techniques offer a simple and well-grounded way to diffuse the effects of the so-called "filter bubble".

Although we investigate only two serendipity-enhancing methods here, additional techniques can easily be introduced to achieve other performance goals. Of particular interest then would be a framework that allows explicit user feedback to shape the algorithm interpolation for individual users, allowing the system to adapt to the adventurousness and mood of different personalities. This could be integrated into a system that maintains serendipity over time, perhaps by cycling through genres or recommendation flavours. Allowing users to direct their own musical growth (through interactive questions or target genres) may also be a way of increasing user satisfaction and promoting musical diversity. Finally, we suggest that a consistent and validated set of performance metrics would greatly aid future work in recommender balance.

Acknowledgments. This work was in part funded by RCUK through the Horizon Digital Economy Research grant (EP/G065802/1). We also thank Òscar Celma for making the dataset publicly available, Stephen Clark for his support, and Toby Moncaster and Jon Crowcroft for their comments. We finally thank the anonymous reviewers.

9. REFERENCES

[1] G. Adomavicius and Y. O. Kwon. Towards more diverse recommendations: Item re-ranking methods for recommender systems. In *Workshop on Information Technologies and Systems*, 2009.

[2] J. An, D. Quercia, and J. Crowcroft. This paper won't change your mind, but...: Why individuals seek diverse opinions (or why they don't). In *Technical Report of the University of Cambridge*, 2012.

[3] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Joural of Machine Learning Research*, 2003.

[4] O. Celma and P. Cano. From hits to niches?: or how popular artists can bias music recommendation and discovery. In *ACM NETFLIX*, 2008.

[5] O. Celma and P. Herrera. A new approach to evaluating novel recommendations. In *ACM RecSys*, 2008.

[6] D. A. Easley and J. M. Kleinberg. *Networks, Crowds, and Markets - Reasoning About a Highly Connected World*. Cambridge University Press, 2010.

[7] T. L. Griffiths and M. Steyvers. Finding scientific topics. *PNAS*, 2004.

[8] J. L. Herlocker, J. A. Konstan, L. G. Terveen, and J. T. Riedl. Evaluating collaborative filtering recommender systems. *ACM Transactions of Information Systems*, 2004.

[9] T. Hofmann. Latent semantic models for collaborative filtering. *ACM Transactions of Information Systems*, 2004.

[10] Y. Hu, Y. Koren, and C. Volinsky. Collaborative filtering for implicit feedback datasets. In *Proceedings of the 2008 Eighth IEEE International Conference on Data Mining*, 2008.

[11] T. Jambor and J. Wang. Optimizing multiple objectives in collaborative filtering. In *ACM RecSys*, 2010.

[12] M. Levy and K. Bosteels. Music recommendation and the long tail. In *WOMRAD*, 2010.

[13] G. Linden, B. Smith, and J. York. Amazon.com recommendations: Item-to-item collaborative filtering. *IEEE Internet Computing*, 2003.

[14] A. K. McCallum. Mallet: A machine learning for language toolkit. `http://mallet.cs.umass.edu`, 2002.

[15] S. M. McNee, J. Riedl, and J. A. Konstan. Being accurate is not enough: how accuracy metrics have hurt recommender systems. In *ACM CHI EA*, 2006.

[16] T. Murakami, K. Mori, and R. Orihara. Metrics for evaluating the serendipity of recommendation lists. In *JSAI*, 2007.

[17] E. Pariser. *The Filter Bubble: What the Internet Is Hiding from You*. Penguin Group USA, 2011.

[18] P. Pu and L. Chen. A user-centric evaluation framework of recommender systems. In *UCERSTI Workshop*, ACM RecSys, 2010.

[19] R. K. Ratner, B. E. Kahn, and D. Kahneman. Choosing less-preferred experiences for the sake of variety. *The Journal of Consumer Research*, 1999.

[20] B. Sarwar, G. Karypis, J. Konstan, and J. Reidl. Item-based collaborative filtering recommendation algorithms. In *ACM WWW*, 2001.

[21] J. B. Schafer, J. Konstan, and J. Riedi. Recommender systems in e-commerce. In *ACM EC*, 1999.

[22] X. Su and T. M. Khoshgoftaar. A survey of collaborative filtering techniques. *Advances in Artificial Intelligence*, 2009.

[23] K. Swearingen and R. Sinha. Beyond Algorithms: An HCI Perspective on Recommender Systems. *SIGIR*, 2001.

[24] K. O. Takayuki Akiyama and M. Tanizaki. Proposal and evaluation of serendipitous recommendation method using general unexpectedness. In *PRSAT*, 2010.

[25] S. Vargas and P. Castells. Rank and relevance in novelty and diversity metrics for recommender systems. ACM RecSys, 2011.

[26] D. J. Watts and S. H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 1998.

[27] T. Zhuo, Z. Kuscik, J.-G. Liu, M. Medo, J. R. Wakeling, and Y.-C. Zhang. Solving the apparent diversity-accuracy dilemma of recommender systems. *PNAS*, 2010.

[28] C.-N. Ziegler, G. Lausen, and S.-T. Lars. Taxonomy-driven computation of product recommendations. In *ACM CIKM*, 2004.

[29] C.-N. Ziegler, S. M. McNee, J. A. Konstan, and G. Lausen. Improving recommendation lists through topic diversification. In *ACM WWW*, 2005.