Measuring Surprise in Recommender Systems

Marius Kaminskas
Insight Centre for Data Analytics
University College Cork, Ireland
marius.kaminskas@insight-centre.org

Derek Bridge
Insight Centre for Data Analytics
University College Cork, Ireland
derek.bridge@insight-centre.org

ABSTRACT

A lot of current research on recommender systems focuses on objectives that go beyond the accuracy of the recommendations; for instance, ensuring that the list of recommended items is diverse. In this work we explore a particular beyond-accuracy objective — serendipity. We follow the idea of measuring item serendipity by its distance from items previously experienced by the user and propose two ways to measure item surprise, which constitutes the core component of serendipitous recommendations.

Through offline experiments we compare three state-of-the-art recommendation algorithms in terms of their ability to generate surprising recommendations. For one of the suggested metrics, the results validate the intuition that a matrix factorization approach generates the most accurate but also the least surprising recommendations, while a user-based neighbourhood approach performs best in terms of surprise. Furthermore, our analysis of the influence of the size of the user's profile on surprise metric values indicates that some information may be lost when averaging item distance values.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Information Filtering*

General Terms

Experimentation, Measurement

Keywords

Recommender systems; evaluation metrics; beyond accuracy; serendipity

1. INTRODUCTION

The goal of recommender systems (RS) research has traditionally been focused on accurately predicting users' ratings for unseen items. However, accuracy is not the only important objective of recommendation [9]. In recent years the focus of RS research has shifted to such objectives as ensuring the recommended items are novel and the set of items is diverse [15]. These objectives are of particular importance since in real life systems users are most likely to consider only a small set of recommendations. It is therefore crucial to make sure that this set of items is as interesting and engaging as possible.

In this paper we investigate serendipity — an objective denoting the ability of a system to generate recommendations that are both surprising and relevant [7]. Although serendipity is frequently mentioned in the research literature as an important complement to recommendation accuracy, until now there have been few attempts to formally define and measure it. A common intuition in the RS literature suggests that, of the state-of-the-art recommendation techniques, matrix factorization approaches and the item-based neighbourhood technique are less likely to make serendipitous recommendations than the user-based neighbourhood technique [5]. However, to the best of our knowledge, this intuition has not been confirmed by experimental evaluation.

While some of the existing approaches to measuring serendipity rely on comparing recommendations against those that would be produced by a baseline recommender system [10, 6], we follow the idea of measuring item surprise by its distance from items previously experienced by the user [11, 1, 16].

In the following section, we review research that addresses the serendipity objective and position our work with respect to the previous efforts. We then propose two alternative ways of measuring surprise, which constitutes the core component of serendipity. Our main contribution lies in comparing the performance of three state-of-the-art recommendation algorithms in terms of their ability to generate serendipitous recommendations and analyzing the impact of user profile size on the surprise metric values. Our findings exemplify the trade-off between accuracy and serendipity, and contribute to the growing body of research on measuring and optimizing recommendation serendipity.

2. RELATED WORK

The term *serendipity*, meaning a "happy and unexpected discovery by accident", was coined in the 18^{th} century [2]. The first studies that recognized the importance of facilitating serendipity in information systems were reported in the information retrieval (IR) literature [14]. In the RS literature, Herlocker et al. [7] informally defined serendipitous

Workshop on 'Recommender Systems Evaluation: Dimensions and Design' (REDD 2014), held in conjunction with RecSys 2014. October 10, 2014, Silicon Valley, United States.

recommendations as "surprisingly interesting items the user might not have otherwise discovered".

2.1 Increasing serendipity

One of the first attempts to optimize a recommender system for serendipity was reported by Iaquinta et al. [8] who considered items with low similarity to a user's profile as potentially serendipitous recommendations and modified a content-based system to include such items alongside the standard recommendations.

Onuma et al. [13] proposed a graph-based algorithm for supporting surprising recommendations. The authors introduced the idea of computing a 'bridge score' of item nodes in the user-item bipartite graph. Nodes connecting separate interconnected areas in the graph receive high bridge scores as they bridge different subspaces in the item information space. The bridge score may be combined with an item relevance score when generating recommendations. Another graph-based approach was proposed by Nakatsuji et al. [11]. The authors applied a Random Walk algorithm on a user similarity graph to identify users that are related (but not too similar) to the target user, arguing that such users provide a good source of surprising recommendations.

Oku and Hattori [12] presented a system that induced possibly serendipitous recommendations by selecting items whose content is a mixture of the content features of two items from the user's profile. Zhang et al. [16] presented a music recommender for Last.fm artists which uses a generative LDA model to build latent clusters of Last.fm users and represent artists with a distribution over these clusters. The authors proposed using the identified latent user communities and artist clustering to generate serendipitous recommendations.

Finally, Adamopoulos and Tuzhilin [1] presented an approach to generate unexpected recommendations by maximizing a utility function which combines an item's relevance and its distance from a set of *expected* items. The set of expected items includes all items rated (i.e., seen) by the user and items similar to the seen ones.

2.2 Measuring serendipity

Although serendipity has been frequently mentioned in the IR and RS literature, few works have provided formal serendipity metric definitions. This is not surprising, as the notion of an item being *surprising* or *unexpected* is difficult to measure without asking the user's opinion — and the notion that the item is *delightful* is probably impossible to capture. It is commonly agreed that serendipity consists of two components – surprise and relevance [7]. While relevance is easy to measure using ratings or other forms of user feedback, surprise (or unexpectedness) of recommendations is difficult to capture without relying on live user studies which are expensive to conduct.

As discussed above, previous approaches to increase serendipity relied on various heuristics to generate more surprising recommendations, e.g., an item being different from a user's profile [8, 1], an item being connected to a distinct area in a user-item graph [13, 11, 16], or an item having a mixture of two input items' features [12]. When evaluating the quality of the results produced by their heuristics, a common practice among the authors is comparing the generated recommendations with recommendations produced by a primitive baseline system (i.e., one which is not optimized for serendipity). This approach to measuring serendipity was proposed by Murakami et al. [10], who argued that a primitive method produces easily predictable items, while the goal of a serendipitous recommender is to suggest items that are difficult to predict. This idea was later adopted by Ge et al. [6], who proposed a formulation of serendipity that combines this notion of unexpectedness with item relevance. This comparative approach to serendipity measurement has a few drawbacks. First, it is sensitive to the choice of the baseline recommender system, and second, since it requires having two recommender systems – one optimized for serendipity and the other not –, it does not allow evaluating the serendipity of existing state-of-the-art recommendation approaches.

Recently, Adamopoulos and Tuzhilin [1] modified the evaluation method of Murakami et al. to one which compares the generated recommendations with a set of *expected* items. Their idea of recommending unexpected items is closely related to our definition of surprise. Similarly to Adamopoulos and Tuzhilin, we also do not require a surprising item to be novel, but only different from the user's expectations, which are represented by a set of items.

The idea of measuring an item's surprise as its distance from a set of *unsurprising* items has been exploited by a few previous works. Nakatsuji et al. [11] defined what they called item *novelty* as the smallest distance from the item class to the class of items previously accessed by user. Vargas and Castells [15], in their framework for measuring diversity and novelty, defined a *personalized novelty* metric based on computing an item's average distance from the user's profile items. The metric was later adopted by Zhang et al. [16].

We follow the idea of Nakatsuji et al. to measure an item's surprise as the *minimum* distance from the user's profile items and we hypothesize that, by contrast, *averaging* the distances between items results in a loss of information, particularly for user's with diverse profiles. Furthermore, unlike some previous works [15, 1], we do not consider item relevance in our metric definitions. While relevance is an important component of serendipity, we leave it to be measured by dedicated accuracy metrics.

3. SURPRISE METRICS

As discussed above, we base our approach to measuring surprise on the intuition that a recommendation is surprising if it is unlike *any* item the user has seen before. We believe this information may be lost when averaging the item distances, especially if the user has been exposed to diverse items. Therefore, we propose using the *lower bound* item distance from the user's profile items as an indicator of surprise. We propose two alternative definitions of surprise, based on different item distance functions.

The first metric exploits users' rating behaviour to measure the likelihood for a pair of items to be seen by the same user. While this information is not direct evidence of item dissimilarity, it provides a reasonable approximation — items that are rarely observed together are likely to be different. The second metric employs a more straightforward item distance function, based on content labels.

Given the target item and the user's profile (a set of items rated by the user), both metrics produce a score that indicates the level of surprise the target item brings to the user. Note that neither metric is presently rank-aware, although they could be adapted to discount items that appear lower in the recommendation list.

3.1 Co-occurrence-based surprise

The first definition is based on the probability for the item to be seen (i.e., rated) together with the items in the user's profile. To measure the pairwise co-occurrence of items we employ normalized point-wise mutual information (PMI) [3], which measures the probability of observing specific outcomes of two independent random variables together. Given a pair of items i and j, we compute their PMI value as:

$$\mathrm{PMI}(i,j) = \log_2 \frac{p(i,j)}{p(i)p(j)} \; / -\log_2 p(i,j) \tag{1}$$

where $p(i)$ and $p(j)$ represent the probabilities for the items to be rated by any user, and $p(i,j)$ is the probability for the same user to rate both items. PMI values range from -1 (in the limit) to 1, with -1 meaning the two items are never rated together, 0 signifying independence of the items, and 1 meaning complete co-occurrence of the items.

In order to measure the surprise of a recommended item i, we compute its PMI with each item in the user's profile, and take the maximum value as the overall surprise value:

$$S_{\text{co-occ}}(i) = \max_{j \in P} \mathrm{PMI}(i,j) \tag{2}$$

where P is the user's profile (i.e., her set of rated items).

Since higher values of $\mathrm{PMI}(i,j)$ signify higher co-occurrence of items i and j (and therefore low surprise of seeing the two items together), taking the maximum value represents the lower bound of the surprise perceived by the user when item i is recommended.

Furthermore, to enable a comparison with the ideas in [15, 1], we conducted experiments with a variant where, instead of relying on the lower bound of surprise, we take the *average* PMI value across items in the user's profile:

$$S^{avg}_{\text{co-occ}}(i) = \frac{\sum_{j \in P} \mathrm{PMI}(i,j)}{|P|} \tag{3}$$

For both $S_{\text{co-occ}}$ and $S^{avg}_{\text{co-occ}}$, lower values signify better surprise results.

Limitations.

We note that the co-occurrence-based definition of surprise metric may be sensitive to rare items, since the point-wise mutual information measure is known to be biased toward rare item pairs [3].

3.2 Content-based surprise

Our second surprise metric is based on distance applied to item content labels. We employed the complement of Jaccard similarity metric for comparing the items:

$$\mathrm{dist}(i,j) = 1 - \frac{L_i \cap L_j}{L_i \cup L_j} \tag{4}$$

where L_i and L_j are sets of labels describing items i and j.

To measure the surprise of a recommended item i, we take the minimum distance between the target item i and the user's profile items:

$$S_{\text{cont}}(i) = \min_{j \in P} \mathrm{dist}(i,j) \tag{5}$$

Taking the minimum distance value as the overall surprise represents the lower bound of how surprising the item is with respect to the seen items.

Similarly to the co-occurrence-based surprise, we also evaluated in our experiment a variant that takes the average distance:

$$S^{avg}_{\text{cont}}(i) = \frac{\sum_{j \in P} \mathrm{dist}(i,j)}{|P|} \tag{6}$$

For both S_{cont} and S^{avg}_{cont}, higher values signify better surprise results.

Limitations.

Note that the content-based metric is sensitive to the quality of item content labels. Low-quality labels may fail to capture the difference between items. Therefore, particular care should be taken when applying the metric to datasets with noisy data, e.g., user-generated labels.

4. EXPERIMENTS

4.1 Datasets

We tested the proposed surprise metrics on two benchmark datasets for offline recommender system evaluation — the MovieLens 1M dataset and LastFM 1K dataset.

The MovieLens dataset contains ∼1 million ratings, 6040 users and 3706 movies. The movies are annotated using a vocabulary of 18 genres (on average 1.65 genres per movie). To obtain richer content descriptors for the movies, we additionally scraped IMDb plot keywords for each movie, which resulted in an average of 81 labels per movie.

The LastFM dataset contains the listening events for 992 users and more than 100K artists. As the dataset is extremely sparse, we cleaned the set of artists by leaving only those for which we could obtain at least three LastFM tags and discarding artists that were listened to by fewer than 20 users. This resulted in 992 users and 7280 artists, with a total of ∼500K ratings. The listening frequencies of the artists were transformed into ratings from 1 to 5 using the standard approach of converting implicit feedback into numerical ratings [15]. Since LastFM content labels for artists are user-generated (and not editorial as in the case of IMDb), to avoid noisy data we retrieved a maximum of the 10 most popular labels for every artist. This resulted in 9.2 labels per artist on average.

We used three state-of-the-art recommendation algorithms in our experiments: *PureSVD* — a matrix factorization approach (with 50 factors) that has been shown to perform well for top-N recommendation tasks [4] — and two standard neighbourhood methods — user-based and item-based collaborative filtering (with a neighbourhood size of 50) [5].

4.2 Evaluation methodology

In recent years, rating-based accuracy metrics for offline RS evaluations have been replaced by precision-oriented metrics that more closely reflect the users' interaction with the system — considering only a small set of top-ranked recommendations, ignoring the lower-ranked items.

In accordance with these state-of-the-art evaluation strategies, in this work we adopt the 'one plus random' methodology [4]. The methodology is based on splitting the dataset into a training set M and probe set P, constructing the test set T by selecting one highly-rated item per user from the probe set. Then, for each user u we make predictions for 1000 unrated items plus the one test item. The set of 1001 items is ranked according to the recommender's predicted score and the top-N recommendations are selected (we used N=10 in our experiments). If the test item is among the top-N items, we have a hit. The overall performance of the system — recall — is calculated as the ratio of the number of hits over the total number of test users. The underlying assumption that the 1000 unseen items are irrelevant is clearly undervaluing the performance, as certain items among the 1000 may be actually relevant for the user.

However, since our work focuses on measuring the surprise of recommendations, we believe this methodology to be an appropriate choice as it involves items the user has not discovered (i.e., unrated items), whereas other offline evaluation strategies employ items the user has had no trouble discovering (i.e., already rated items).

As described in Section 3, given an item and the user's profile, the co-occurrence-based and content-based surprise metrics produce a score in $[-1;1]$ and $[0;1]$ respectively. To obtain a single surprise value for a user's top-N recommendation list, we average the surprise values of all recommended items.

We computed the results using 5-fold cross validation with 80%-20% training/probe set split.

5. RESULTS AND DISCUSSION

As a first step in the evaluation, to make sure that the two surprise metrics are not redundant, we compared them using the following procedure: for each user in the dataset, we generated a list of top-10 recommendations using the *PureSVD* matrix factorization approach. We then ranked the 10 items using each metric and computed Spearman's rank correlation for the two rankings. Averaging the results over all users we obtained values close to zero for both MovieLens and LastFM datasets. This indicates that the two metric definitions capture different aspects of surprise.

In accordance with this finding, the two surprise metrics give different results in our main experiment.

5.1 Surprise value comparison

Table 1 lists the average recall and surprise values across test users for the evaluated algorithms — matrix factorization approach (MF), user-based approach (UB), and item-based approach (IB) — on both datasets. All reported metrics produce significantly different means when applied to different algorithms ($p < 0.001$ in a one-way ANOVA test). To measure which algorithms perform better/worse than the others, we applied the two-tailed t-test to each pair of algorithms.

Table 1: Recall and surprise results for the three tested algorithms. For each metric, values marked with $^{+}/^{-}$ are significantly better/worse than the other two values with $p < 0.001$.

	MovieLens 1M			LastFM 1K		
Metric	MF	UB	IB	MF	UB	IB
Recall	0.334^{+}	0.022^{-}	0.065	0.427^{+}	0.051	0.059
$S_{\text{co-occ}}$	0.349	0.349	0.354^{-}	0.379^{+}	0.404	0.407^{-}
S_{cont}	0.915^{-}	0.928^{+}	0.927	0.480^{-}	0.643^{+}	0.613
$S^{avg}_{\text{co-occ}}$	0.177^{-}	0.145	0.094^{+}	0.141^{-}	0.012^{+}	0.050
S^{avg}_{cont}	0.976^{-}	0.981^{+}	0.980	0.887^{-}	0.938^{+}	0.927

As expected, MF produces the most accurate recommendations, outperforming both neighbourhood approaches.

For the content-based surprise metric (where higher values are better), the results are consistent for both lower-bound distance (Eq. 5) and average distance (Eq. 6) metric variants — MF performs significantly worse than both UB and IB, while UB performs better than IB.

On the other hand, for the co-occurrence-based surprise metric (where lower values are better), results differ for the lower-bpund distance (Eq. 2) and average distance (Eq. 3) variants. In case of the lower-bound distance $S_{\text{co-occ}}$ variant, the MF approach performs as well as the UB and IB approaches, even outperforming both of them on the LastFM dataset. Contrastingly, the average distance variant $S^{avg}_{\text{co-occ}}$ shows MF to be outperformed by both UB and IB techniques.

An explanation for results of the co-occurrence-based surprise metric may lie in its sensitivity to rare items. In our experiments, the matrix factorization approach tends to rank popular items higher than both neighbourhood approaches, which can influence the values of point-wise mutual information. When one of the items in $\text{PMI}(i,j)$ is rare, the normalized PMI metric can take extreme values — close to 1 if the two items frequently co-occur in the dataset, or close to -1 if they almost never occur together. Consequently, the lower-bound distance metric variant (which takes the maximum PMI value of an item's distance to a user's profile) tends to produce higher values when applied to rare items, while the average distance metric variant tends to produce low overall scores as it gets reduced by the negative PMI values. More experiments are needed to confirm this intuition. If confirmed, the $S_{\text{co-occ}}$ metric could be modified to avoid a bias toward rare items [3].

5.2 The impact of user's profile size

The size of a user's profile is an important factor to consider when measuring surprise of recommendations. Intuitively, the more items a user has been exposed to, the more difficult it is to surprise the user. To test this intuition, we plotted the surprise values obtained for individual users against their profile size. For both surprise metrics, we plotted two variants of results — one obtained using the lower-bound distance from the user's profile, and the other using the average distance from the profile. In both cases we measured the surprise of the recommendations generated using the most accurate recommendation method — the *PureSVD* approach.

Results for the *lower-bound* distance metric variant for both datasets are shown in the left half of Figure 1. As the user profile size increases, surprise values for the $S_{\text{co-occ}}$ metric tend to increase and the values for the S_{cont} metric decrease, i.e., in both cases the surprise of recommendations is decreasing. This finding confirms the intuition that given a large user's profile, there is a higher chance of finding an item similar to the recommended item, therefore, the chance of surprising the user is smaller.

The same trend is not observed when computing surprise using the *average* item's distance from a user's profile (right half of Figure 1), as the surprise values are independent of the profile size. We therefore believe that the lower-bound distance is more appropriate when measuring an item's surprise with respect to a user's profile, since averaging item distances results in losing information about the user's past

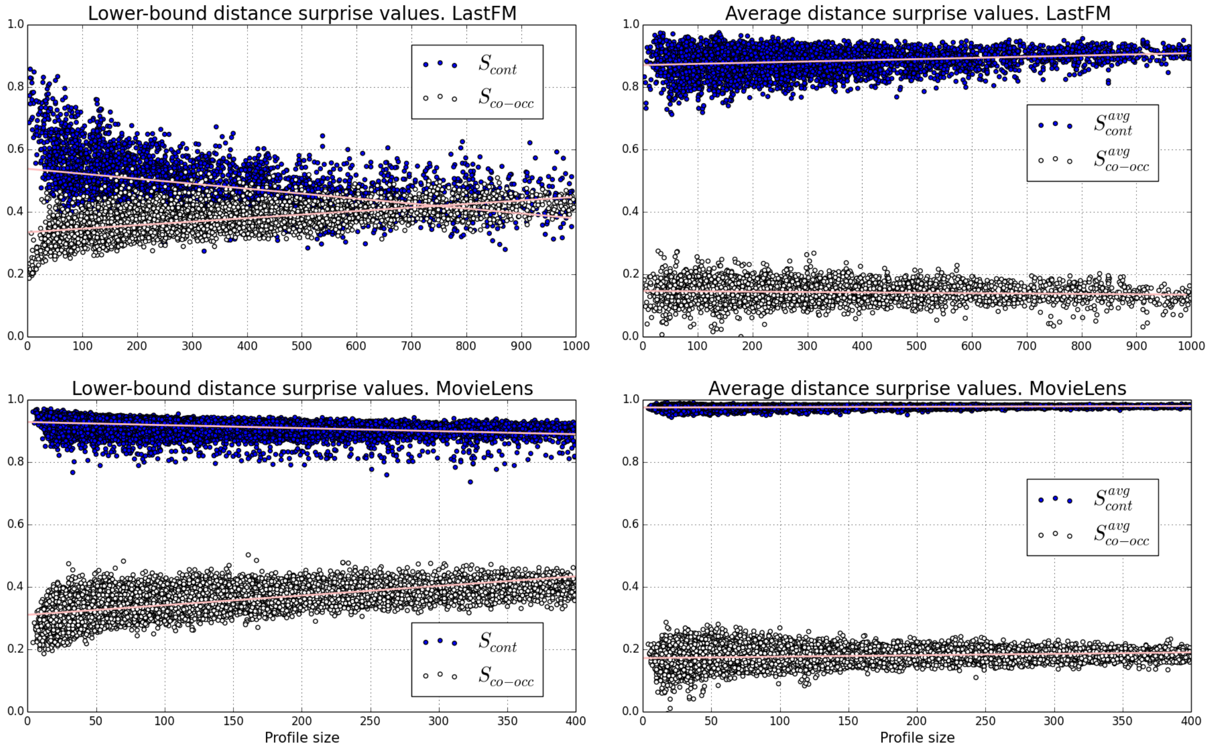

Figure 1: Profile size influence on surprise values.

interactions with the system.

5.3 Qualitative analysis

For a qualitative insight into the results, we looked at the recommendations generated for individual users. Table 2 shows for one user from the LastFM dataset 10 items from her profile and the top-10 recommendations generated using each algorithm. A subset of content labels is displayed for each item. (The full set of labels is omitted due to space limitations.) As can be seen from the content labels, all three algorithms generate items that are related to the user's musical preferences — independent rock and electronic music. However, user-based recommendations include more artists that may be surprising discoveries for the user, e.g., a jazz musician and a string quartet playing covers of popular songs in classical style. The item-based approach generates one such recommendation — a heavy metal band playing rockabilly songs.

6. CONCLUSIONS AND FUTURE WORK

In this paper, we investigated two metrics – co-occurrence-based ($S_{\text{co-occ}}$) and content-based (S_{cont}) – for measuring surprise in recommender systems. We evaluated each metric with three state-of-the-art recommenders on two benchmark datasets. For both co-occurrence- and content-based metrics, two variants were evaluated — one using the lower-bound item's distance from the user's profile items as an indicator of surprise, and the other using the average distance from the profile items.

Using the S_{cont} surprise metric, the experimental results confirm that there is a trade-off between recommendation accuracy and serendipity [6] and the intuition that the matrix factorization approach, while being the most accurate, produces the least surprising recommendations with respect to the user's profile. Conversely, the user-to-user collaborative filtering approach produces the most surprising recommendations.

The above finding is confirmed for both S_{cont} metric variants — the lower-bound distance and average distance. However, the analysis of user's profile size influence on surprise values indicates that some information may be lost when using the average distance metric variant. Therefore, we believe that using the lower-bound distance to profile items is the better approach for measuring surprise.

The findings of S_{cont} are not fully confirmed using the alternative, co-occurrence-based surprise metric $S_{\text{co-occ}}$. We believe the reason for this lies in the metric's sensitivity to rare items. However, the metric deserves further investigation, since in contrast to the content-based metric, it does not rely on item content labels and may capture a different aspect of surprise than the content-based metric.

The next steps in our research include comparing the proposed metrics against existing serendipity metrics. Furthermore, we are interested in measuring how the state-of-the-art recommendation algorithms perform with respect to other beyond-accuracy objectives, such as diversity, novelty, and coverage, as well as investigating correlations and trade-offs between these objectives and serendipity.

Most importantly, a user study is needed to validate the effectiveness of the proposed metrics. While offline evaluations have demonstrated the trade-off between recommendation accuracy and surprise, obtaining user opinions in a live study is essential to understanding which surprising recommendations are appreciated by the users.

Table 2: A sample user profile and the generated recommendations from the LastFM dataset. The light-shaded item is the recovered hidden test item. The dark-shaded items are musicians belonging to music genres not present in the user's profile — jazz, classical, and heavy metal music.

Profile	PureSVD (MF)	$s_{\text{co-occ}}$	s_{cont}	User-based (UB)	$s_{\text{co-occ}}$	s_{cont}	Item-based (IB)	$s_{\text{co-occ}}$	s_{cont}
Death Cab for Cutie indie rock, alternative, emo, pop	Bloc Party indie rock, alternative, britpop	0.318	0.333	Liam Finn indie, singer-songwriter, folk, pop	0.357	0.750	Gram Parsons country, singer-songwriter, folk, rock	0.497	0.571
John Mayer singer-songwriter, acoustic, rock, blues	The Smiths indie, new wave, alternative, britpop	0.381	0.333	Quietdrive pop, punk, rock, alternative, emo	0.395	0.571	Ralph Myerz electronic, downtempo, chillout, funk	0.390	0.667
Damien Rice singer-songwriter, acoustic, indie, folk	Metric indie rock, female, alternative	0.379	0.461	Jan Garbarek jazz, saxophone, instrumental	0.446	0.889	To My Boy new rave, electronic, indie, power pop	0.365	0.750
Travis britpop, indie rock, alternative	M83 electronic, shoegaze, post-rock, indie	0.377	0.667	Steve Aoki electronic, techno, indie, dance	0.379	0.571	The Enemy indie rock, alternative, britpop, punk	0.466	0.333
The New Pornographers indie rock, alternative, power pop	Eels indie, alternative, rock, pop, american	0.326	0.333	Girl Talk electronic, dance, hip-hop, mashup	0.347	0.667	Volbeat heavy metal, rockabilly, rock and roll	0.321	0.750
The Postal Service indie, electronic, pop, alternative, emo	Pink Floyd rock, progressive, psychedelic, classic	0.264	0.571	Bloc Party indie rock, alternative, britpop	0.318	0.333	Seasick Steve blues, rock, folk, country	0.458	0.571
Toby Keith modern country, male, pop, american	Ratatat electronic, instrumental, indie	0.337	0.462	Ralph Myerz electronic, downtempo, chillout, funk	0.390	0.667	Elf Power elephant 6, indie rock, lo-fi	0.422	0.333
The Hold Steady indie rock, alternative, post-punk	Green Day punk rock, alternative, pop punk	0.246	0.571	Olivia Ruiz french, chanson francaise, female	0.394	0.750	Bloc Party indie rock, alternative, britpop	0.318	0.333
Clap Your Hands Say Yeah indie rock, alternative, indie pop, lo-fi	Girl Talk electronic, dance, hip-hop, mashup	0.347	0.667	Love And Rockets post-punk, new wave, gothic, 80s	0.355	0.182	Biffy Clyro alternative, indie rock, emo	0.341	0.462
The View indie rock, britpop, garage rock, punk	MSTRKRFT electronic, dance, house, indie	0.329	0.571	The String Quartet instrumental, classical, covers, strings	0.258	0.889	Boys Night Out emo, screamo, rock, hardcore	0.345	0.571
Average values:		0.330	0.497		0.364	0.627		0.392	0.534

7. ACKNOWLEDGMENTS

This publication has emanated from research conducted with the financial support of Science Foundation Ireland (SFI) under Grant Number SFI/12/RC/2289.

8. REFERENCES

[1] P. Adamopoulos and A. Tuzhilin. On Unexpectedness in Recommender Systems: Or How to Better Expect the Unexpected. *Working Paper: CBA-13-03, New York University*, 2013.

[2] P. André, M. Schraefel, J. Teevan, and S. T. Dumais. Discovery is never by chance: designing for (un)serendipity. In *Procs. of the 7th ACM Conference on Creativity and Cognition*, pages 305–314, 2009.

[3] G. Bouma. Normalized (pointwise) mutual information in collocation extraction. In *Procs. of the Conference of the German Society for Computational Linguistics and Language Technology*, pages 31–40, 2009.

[4] P. Cremonesi, Y. Koren, and R. Turrin. Performance of recommender algorithms on top-n recommendation tasks. In *Procs. of the 4th ACM Conference on Recommender Systems*, pages 39–46, 2010.

[5] C. Desrosiers and G. Karypis. A comprehensive survey of neighborhood-based recommendation methods. In F. Ricci et al., editors, *Recommender Systems Handbook*, pages 107–144. Springer, 2011.

[6] M. Ge, C. Delgado-Battenfeld, and D. Jannach. Beyond accuracy: evaluating recommender systems by coverage and serendipity. In *Procs. of the 4th ACM Conference on Recommender Systems*, pages 257–260, 2010.

[7] J. L. Herlocker, J. A. Konstan, L. G. Terveen, and J. T. Riedl. Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems*, 22(1):5–53, 2004.

[8] L. Iaquinta, M. de Gemmis, P. Lops, G. Semeraro, M. Filannino, and P. Molino. Introducing serendipity in a content-based recommender system. In *Procs. of the 8th Conference on Hybrid Intelligent Systems*, pages 168–173, 2008.

[9] S. M. McNee, J. Riedl, and J. A. Konstan. Being accurate is not enough: how accuracy metrics have hurt recommender systems. In *Extended Abstracts on Human Factors in Computing Systems*, pages 1097–1101, 2006.

[10] T. Murakami, K. Mori, and R. Orihara. Metrics for evaluating the serendipity of recommendation lists. In *Procs. of the 2007 Conference on New Frontiers in Artificial Intelligence*, pages 40–46, 2008.

[11] M. Nakatsuji, Y. Fujiwara, A. Tanaka, T. Uchiyama, K. Fujimura, and T. Ishida. Classical music for rock fans?: Novel recommendations for expanding user interests. In *Proceedings of the 19th ACM Conference on Information and knowledge management*, pages 949–958, 2010.

[12] K. Oku and F. Hattori. Fusion-based recommender system for improving serendipity. In *Procs. of the Workshop on Novelty and Diversity in Recommender Systems (DiveRS 2011)*, pages 19–25, 2011.

[13] K. Onuma, H. Tong, and C. Faloutsos. Tangent: a novel, 'surprise me', recommendation algorithm. In *Procs. of the 15th ACM Conference on Knowledge discovery and data mining*, pages 657–666, 2009.

[14] E. G. Toms. Serendipitous information retrieval. In *Procs. of the 1st Workshop on Information Seeking, Searching and Querying in Digital Libraries*, pages 11–14, 2000.

[15] S. Vargas and P. Castells. Rank and relevance in novelty and diversity metrics for recommender systems. In *Procs. of the 5th ACM Conference on Recommender Systems*, pages 109–116, 2011.

[16] Y. C. Zhang, D. Ó. Séaghdha, D. Quercia, and T. Jambor. Auralist: introducing serendipity into music recommendation. In *Procs. of the 5th ACM Conference on Web search and data mining*, pages 13–22, 2012.