Understanding how to Explain Package Recommendations in the Clothes Domain

Agung Toto Wibowo
Computing Science / Informatics Engineering
University of Aberdeen / Telkom University
wibowo.agung@abdn.ac.uk /
agungtoto@telkomuniversity.ac.id

Advaith Siddharthan
Knowledge Media Institute
The Open University
advaith.siddharthan@open.ac.uk

Judith Masthoff
Computing Science / Information and Computing Science
University of Aberdeen / Utrecht University
j.f.m.masthoff@uu.nl/j.masthoff@abdn.ac.uk

Chenghua Lin
Computing Science
University of Aberdeen
chenghua.lin@abdn.ac.uk

ABSTRACT

Recommender system explanations have been widely studied in the context of recommending individual items to users. In this paper, we present and evaluate explanations for the more complex problem of package recommendation, where a combination of items that go well together are recommended as a package. We report the results of an empirical user study where participants try to select the most appropriate combination of a "top" (e.g. a shirt) and a "bottom" (e.g. a pair of trousers) for a hypothetical user based on one of 5 types of explanation communicating item-feature preferences and/or appropriateness of feature combinations. We found that the type of explanation significantly impacted decision time and resulted in selection of different packages, but found no difference in how participants appraised the different explanation types.

KEYWORDS

Package Explanation; Package Recommendation; Clothes Domain

1 INTRODUCTION

Recommender systems have been widely studied in various domains (e.g. movies, music, books) using techniques such as collaborative filtering [3, 10], content based filtering [9], and hybrid methods [11]. These techniques have been used in different tasks such as finding good items (top-N recommendations) [2], recommending a sequence [1, 7] or recommending a package (also referred to as a bundle) [4, 8].

To increase user acceptance of recommended items, recommender systems can provide explanations [5, 14], which also help users understand why items are being selected for them [6]. Explanations can provide transparency by exposing the reasoning and data behind a recommendation [5, 13], and increase user scrutability, trust, effectiveness, persuasiveness, efficiency and satisfaction [13, 14].

Several explanations types have been investigated [14]. Explanations have used different inputs such as content data and user-item ratings [6]. In the movie domain, a comprehensive study has been conducted by presenting different explanation interfaces [5] such as using group rating histograms, neighbor ratings histograms or tables of neighbor ratings. Using user-item ratings as input, Hernando et al. [6] proposed tree explanations where the nodes represent items, and the branches represent the distance among items. Explanations can also take the form of social explanations. This type of explanation is usually delivered in the form of "users u_1 and u_2 also like the recommended item" [16].

Other explanations use items features. For example, in the movie domain explanations can use features such as the director and actors [12], or use tags (free text describing an item) as studied in tagsplanations [15]. In tagsplanations, tags were presented along with the item relevances and user preferences. The tag relevance may reflect the tags' popularity, or the correlation between the tags and items, whilst user preferences measure users' sentiment to the given tag, e.g. how much a user will like or dislike the "classic" tag.

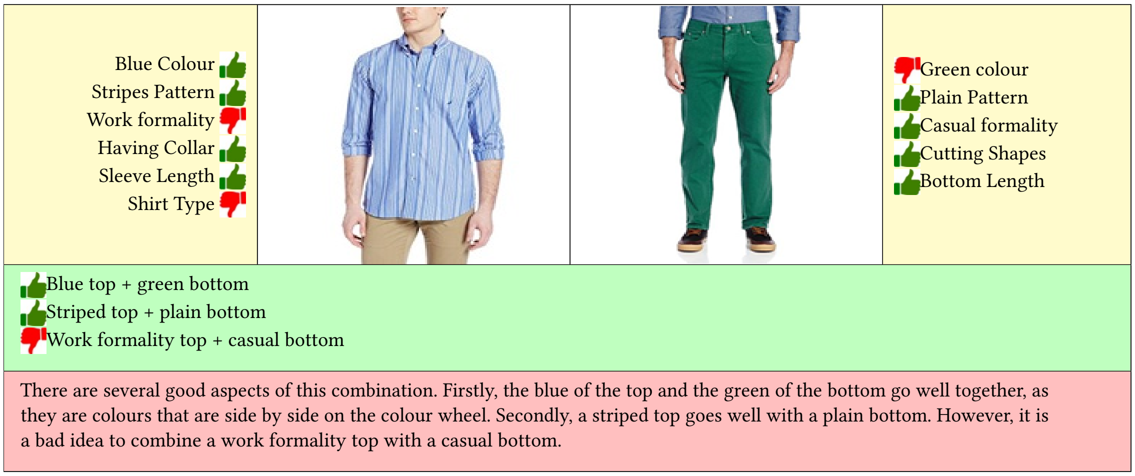
All the explanations discussed above were used to explain recommendations of *individual* items [14]. Recently, a more complex task has been studied in the form of package recommendations [4, 17, 18], where combinations of items that work well together are recommended to a user. In real world applications, package recommendations have advantages for both customers and sellers. For example, in the clothes domain, when a "top"(e.g. a shirt) and "bottom" (e.g. a pair of trousers) are recommended together, the seller can boost their sales, while the customer, they can save on shipping costs and obtain clothes that go well together.

To the best of our knowledge there is a lack of research on explanations which deal with package recommendations. This type of explanation has two main challenges: it must explain both the individual items in the combination and the appropriateness of combining them. In this paper, within the clothes domain, we investigate the impact of five different types of explanation by combining three different components, namely, individual preferences, package appropriateness, and natural language descriptions of package appropriateness.

The remainder of this paper is organized as follows. Section 2 defines the package recommendation and explanation in clothes domain. Section 3 describes our motivation, participants, materials. Section 4 shows our results. Finally, Section 5 provides a discussion and suggests directions for future work.

2 CLOTHES PACKAGE EXPLANATIONS

In this paper, we followed package definitions as described in [19]. Consider a set of clothes consisting of two disjoint complementary sets: a set of "top" items and a set of "bottom" items. Each item in both "top" and "bottom" is associated with a set of attributes (for



There are several good aspects of this combination. Firstly, the blue of the top and the green of the bottom go well together, as they are colours that are side by side on the colour wheel. Secondly, a striped top goes well with a plain bottom. However, it is a bad idea to combine a work formality top with a casual bottom.

Figure 1: Clothes explanation components for a combination of "top" and "bottom" of clothes (white cells). (a). Explanation using top and bottom individual thumb up/down attributes (yellow cells, used in IT and IT-CT explanation types/see Table 1), (b). Explanation using top and bottom relations using thumb up/down (green cell, used in CT and IT-CT), and (c). Explanation using top and bottom relations using natural language description (red cell, used in CN and IT-CN).

Table 1: Clothes Combination Explanation Types

Explanation Type	Indiv. Thumb	Combination Thumb	Nat. Lang.
Indiv. Thumb (IT)	✓	-	-
Comb. Thumb (CT)	-	✓	-
Comb. Nat. Lang. (CN)	-	-	✓
Indiv. Thumb + Comb. Thumb (IT-CT)	✓	✓	-
Indiv. Thumb + Comb. Nat. Lang. (IT-CN)	✓	-	✓

example colour, pattern, formality and so on). Further, some of these items and/or their combinations (a package) have received ratings from one (or more) users as individual rating and/or package rating. Our task is then to provide explanations for selected packages.

To highlight the importance of package explanations, consider a situation when a user looking a complementary item for a t-shirt (as a query) he/she like. The recommender engine might pair the user's query with other complementary items as packages. These packages will be better served with explanations. The explanation also useful in a situation where a user request different packages.

In the literature, different explanations have used different intermediary entities to show the relations between users and items [15]. Three commonly used intermediary entities are items, users, and features. In this paper, we use features (see Figure 1), as they are easier to detect and explain.

Using features, we designed three different explanation components (see Figure 1). The first component (in the yellow boxes) uses individual attributes for both the "top" and the "bottom" clothes. We use colour, pattern, formality, collar, sleeve length and type (e.g. shirt, t-shirt or top) as features for the top, and colour, pattern, formality, cutting shapes, and length as features for the bottom. In clothes domain we can easily extract and communicate these attributes by seeing it from each image. The thumb up/down symbols indicate the user preferences (like/dislike) of the feature values. Recommender system might calculate the user preferences by correlating each attribute with the user individual ratings. The second component (in the green box) uses the relation between the top and bottom clothes in the form of appropriateness rules which we adopted from [19], and presents this relation using thumb up/down symbols. In this component, the thumb up/down represents the appropriateness/ inappropriateness of attributes from the top and bottom being combined together. In the real world situation, a user might have an intuition and easily judge whether this component correctly explain the relation among "top" and "bottom" or not. Following [19], we use color, pattern and formality as combination features. The third component (in the red box) uses the same appropriateness rules as the second component and manually describes the top and bottom combination in natural language. We used the natural language to reduce miss-interpretation to the provided explanations. All three explanation components can be system generated, but for this study were manually generated to ensure our findings about how users appraise these components were not influenced by issues pertaining to implementation quality.

Using these components, we designed five different explanation types (see Table 1). We named the explanations using the abbreviation of components involved in each type. For example, IT-CN is the explanation which uses the individual thumb (yellow cells in Figure 1) and the combination natural language components (red cell in Figure 1). In this study, we did not use CT and CN components together as they present similar information.

3 EVALUATION METHODOLOGY

The aim of our user study was to evaluate different package recommendation explanation types (see Table 1) in the clothes domain on different aspects (such as effectiveness/persuasiveness, efficiency, transparency, trust, and satisfaction).

Table 2: Number of images shown in the preferences sheet.

Pseudo-user	# of Top Ratings					# of Bottom Ratings				
	1	2	3	4	5	1	2	3	4	5
Mary	2	4	3	4	3	2	3	2	2	3
Peter	3	3	2	4	3	3	2	1	3	3

3.1 Participants

In our study, 64 participants were recruited using convenience sampling at University of Aberdeen. Participants came from 17 different countries (2 did not disclose their nationality). There were 38 male participants, 25 female, and 1 undisclosed. There were 10 participants aged between 18-25, 45 between 26-45, 3 between 41-54, and 1 undisclosed. The study took place in an office environment, and was approved by the University's Ethics Board.

3.2 Materials

We created two pseudo-users, named "Mary" and "Peter", using real data from [19]. We showed the participant some images of "top" and "bottom" clothes and their real ratings by the pseudo-user. We selected the clothes at random, and slightly varied the number of clothes shown (see Table 2 for the number of images shown for each rating, e.g., the "3" in the far bottom right cell indicates the number of "bottom" images rated "5" by Peter shown to participants).

We selected six combinations of tops and bottoms with the explanations that the system would have generated for the pseudo-user for each explanation type (see Table 1). Table 3 shows for each combination how many positive and negative aspects were mentioned in the explanations for the "top", "bottom" and "combination" respectively. For example, Figure 1 shows the explanation components for Peter's combination C3 which contains 4 positive attributes for the "top" and the "bottom" each, and also 2 positive appropriateness aspects for the "combination".

3.3 Experimental Design

We used a mixed design. Each participant considered the two pseudo-users, with a different explanation type for each pseudo-user. The explanation type used for the pseudo-users was counterbalanced, as was the order in which pseudo-users were considered. Four groups of 16 participants each considered different pairs of explanation types: (1) CT and CN, (2) IT and IT-CT, (3) IT and IT-CN, and (4) IT-CT and IT-CN. This enabled a within subject comparison between the explanation types in each pair, but also a between-subject comparison for other pairs (such as CT versus IT-CT).

Independent variables:

- Explanation type: 5 types (IT, CT, CN, IT-CN, IT-CT), which varied in the *Explanation of a package's individual items* (2 values: included in graphical thumbs up/down form or not included), and the *Explanation of the package's combination aspects* (3 values, included in graphical thumbs up/down form, included in natural language form, or not included in either form). We excluded the explanation type without any explanations of either item or combination.
- Pseudo-user: two values, Mary and Peter.

Table 3: Positive and negative aspects' frequency distribution in each combination

Pseudo-user	Comb #	Top		Bottom		Combination	
		Pos.	Neg.	Pos.	Neg.	Pos.	Neg.
Mary	C1	6	0	3	2	2	1
	C2	5	1	4	1	3	0
	C3	4	2	5	0	2	1
	C4	4	2	4	1	2	1
	C5	6	0	3	2	3	0
	C6	5	1	4	1	3	0
Peter	C1	4	2	4	1	2	1
	C2	3	3	5	0	3	0
	C3	4	2	4	1	2	1
	C4	4	2	4	1	2	1
	C5	3	3	5	0	3	0
	C6	4	2	4	1	2	1

Dependent variables:

- Five perceived qualities of an individual explanation type, each rated on a 7-point scale (see Procedure for details): (1) Effectiveness, (2) Trust, (3) Efficiency, (4) Transparency, (5) Overall quality of the explanation type.
- Actual efficiency: Speed of package selection t in seconds.
- Package selected.
- Comparative preference for two explanation types.

3.4 Procedure

Participants were told that the purpose of the study was to understand the effectiveness of different explanations about clothes combinations. After providing informed consent, the study was run using the following steps:

Step 1. Participants provided demographic information (gender, age and nationality, with the option not to disclose).

Step 2. Participants were given a pseudo-user's preferences about individual clothing items (see Section 3.2 and Table 2).

Step 3. Participants selected a combination of clothes for this user out of 6 combinations provided, with all combinations including an explanation of the same style (see Table 1; different users saw different explanation styles). The decision time was recorded.

Step 4. Participants answered six questions regarding the decision process and the explanations (Question 1 on a scale of 1 to 5, the others from 1 to 7):

(1) What rating do you think the user will give to the combination you have chosen? (This question was only asked to enable posing the next question.)
(2) How confident are you that your rating reflects the rating the user would have given? (This question was used to measure the impact of explanation style on effectiveness – whether the explanations help users make good decisions [Efk.][1].)
(3) Please rate how easy it is to decide how good combinations would be for the user? (This question was used to measure the impact of explanation style on perceived efficiency [Efc.]).

[1] One can argue whether this is in fact effectiveness or persuasiveness, as it only measures the extent to which the participant thinks the user will agree with them, which does not necessarily make the rating correct.

Table 4: Statistical comparatives between two explanation types

		Mean (StDev) for Expl. Type 1						Mean (StDev) Expl. Type 2						Pref.**
Type 1	Type 2	t (s)	Efk.	Efc.	Tra.	Trust	Sat	t (s)	Efk.	Efc.	Tra.	Trust	Sat.	
CT	CN	77.4 (39.2)*	6.1 (0.9)	5.1 (1.4)	5.6 (1.2)	4.9 (1.1)	5.2 (1.6)	123.4 (55.3)*	5.8 (1.1)	4.9 (1.8)	5.4 (1.5)	4.9 (1.3)	5.0 (1.7)	4.0 (2.5)
IT	IT-CT	114.6 (75.3)	5.6 (1.1)	4.1 (1.2)	5.3 (1.0)	5.1 (0.9)	5.4 (1.0)	140.1 (94.0)	5.3 (1.1)	4.2 (1.6)	5.4 (1.0)	4.9 (1.1)	5.5 (1.2)	3.9 (2.3)
IT	IT-CN	91.6 (62.7)*	6.1 (0.9)	5.6 (1.1)	5.6 (0.9)	5.4 (0.6)	5.9 (0.9)	137.7 (69.5)*	5.7 (1.0)	5.1 (1.5)	5.8 (0.9)	5.1 (0.9)	6.0 (1.0)	4.9 (2.1)
IT-CT	IT-CN	85.9 (42.2)*	5.5 (1.1)	4.4 (2.2)	5.0 (1.5)	4.9 (1.6)	5.6 (1.5)	146.9 (73.6)*	5.9 (0.7)	4.4 (1.6)	5.2 (1.2)	5.1 (1.1)	5.9 (0.9)	4.0 (2.4)

Paired T-test, * significant at $p < 0.05$. ** User preferences when comparing explanation types from 1 (strongly prefered type 1) to 7 (strongly prefered type 2).

(4) The system will in future recommend clothing combinations. To what extent do the explanations make you understand what the system will base its recommendation for a user on? (This question was used to measure the impact of different explanation types on perceived transparency [Tra.]).

(5) To what extent would you trust the system to produce recommendation of clothing combinations for a user? (This question was used to measure the impact of different explanation types on user trust).

(6) Overall, how much do you like the explanations provided? (This question was used to measure satisfaction [Sat]).

Step 5 Steps 2-4 were repeated for the other pseudo-user with a different explanation style.

Step 6 Participants rated on a scale of 1-7 their relative preference for the two explanation styles they had seen (with 1 meaning they strongly preferred one particular style, and 7 meaning they completely preferred the other style). Participants were also asked to provide additional comment (optional) regarding the two explanation styles they had seen.

4 RESULTS

Do explanation types impact perceived quality? Table 4 shows participants' ratings of perceived effectiveness (Efk.) efficiency (Efc.), transparency (Tra.), trust and satisfaction (Sat.), as well as their explicit comparative preference rating. No significant impact of the explanation type on these perceived quality measures was found, and participants overall seemed to equally appreciate all explanation types (with participants being clearly satisfied with all)[2]. A post-hoc analysis showed that individual participants often did prefer one of the two explanations (and also gave higher perceived quality ratings for that one), but varied in which type they preferred. So, overall, we did not find an impact of explanation type on perceived quality, but wonder whether the choice of explanation type may need to be adapted to the individual user.

Do explanation types impact decision speed? Table 4 shows the time (t) participants took to make their combination selection. Participants were significantly faster with CT than CN, with IT-CT than IT-CN, and with IT than IT-CN, implying that reading the natural language explanations slowed them down. However, there was no significant difference between IT and IT-CT (whilst there may seem to be trend for IT to be faster, this does not hold up when combining the data from other rounds in which IT and IT-CT were used). So, adding the combination component did not actually make participants slower, as long as it used the thumbs up/down rather than natural language. Combining the data from the multiple rounds, we also find that participants were significantly faster with CT than IT-CT (t-test, p<0.01), so adding the individual explanation component slowed them down. Overall, participants were fastest with CT. So, explanation type clearly impacts decision speed and from an actual efficiency point of view, CT performed best.

Table 5: Percentage distribution of selected combination per explanation type

Pseudo-user	Explanation Type	C1	C2	C3	C4	C5	C6
Mary	IT	0%	13%	81%	0%	0%	6%
	CT	13%	0%	50%	25%	0%	13%
	CN	0%	13%	75%	0%	0%	13%
	IT-CT	0%	19%	56%	6%	13%	6%
	IT-CN	6%	38%	38%	13%	0%	6%
Peter	IT	0%	6%	19%	31%	19%	25%
	CT	0%	13%	0%	50%	25%	13%
	CN	0%	13%	25%	38%	13%	13%
	IT-CT	0%	19%	13%	19%	13%	38%
	IT-CN	0%	50%	6%	19%	13%	13%

Do explanation types impact decisions made? Table 5 shows the percentage distribution of the selected combination (see Table 5) per explanation type. The green cells show the most selected combination. Interestingly, when we showed participants only one explanation component (IT, CT, or CN), the most selected combination was the same independent of explanation type. However, when we showed both the individual and combination components (IT-CT or IT-CN), participants' selections tended to change. For example, participants who received the IT explanation type for Mary tended to select C3. The distribution changed to C2 when participants received IT-CN and changed slightly towards C2 and C5 when participants received IT-CT. Both C2 and C5 were the combinations with 3 positive combination aspects, whilst C3 was a combination with 2 positive and 1 negative combination aspects (see Table 3). For Peter, the change is even more pronounced, with a change from C4 towards C6 and C2. A posthoc Chi-square test of independence was performed to examine the relation between explanation type (one component or two components) and the combination selected (only considering those combinations which were chosen most often, namely C2, C4 and C6 for Peter, and C2 and C3 for Mary). This test was statistically significant for both pseudo-users (p<0.05). So, the explanation type impacts the decisions people make, implying that explanations may either influence the effectiveness or persuasiveness of package recommendations.

[2] There was a significant effect with participants preferring IT-CN to IT (z-test shows the mean to be significantly above 4, p=0.04, however the perceived quality metrics are not significant and the trend on several is in the opposite direction).

Participants' additional feedbacks. At the end of our survey, we asked our participants to provide additional comment. This question was optional and we received two valuable issues for future improvements. First, the explanations need to be presented in more detail. For example, in IT, the sleeve length information is better served with short/long sleeve. Second, even though the explanation types which involved CN took longer time in the package selection process, however, some participants said this type of explanation is useful to help them decide the selection process. This was expressed from a comment such as: "*explanation may be useful, however, too long to understand*". To handle this problem different participant suggested to deliver CN component in different ways: "*explanation should be more explicit, may be with bold letter for points that are stressed or colours*".

5 CONCLUSION AND FUTURE WORK

This paper provides early insights into explanations for package recommendations. The type of explanation had a significant impact on decision time, but no difference in perceived quality was found (though there was some evidence that different people preferred different explanation types). We also found that different explanation types differently impacted the selection of packages.

This study was conducted in the clothes domain, so its generalisability in other domains such as travel services needs to be investigated. The study can also be extended by expanding the current package explanations to incorporate items the user liked (or disliked). We would also like to study package explanations more directly in a real world situation where the explanations are for real users for whom the system makes recommendations.

ACKNOWLEDGMENTS

We would like to thank Lembaga Pengelola Dana Pendidikan (LPDP), Departemen Keuangan Indonesia for awarding a scholarship to support the studies of the lead author. We would also like to thank the participants who provided precious feedback during our study.

REFERENCES

[1] Da Cao, Liqiang Nie, Xiangnan He, Xiaochi Wei, Shunzhi Zhu, and Tat-Seng Chua. 2017. Embedding Factorization Models for Jointly Recommending Items and User Generated Lists. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval.* ACM, 585–594.

[2] Paolo Cremonesi, Yehuda Koren, and Roberto Turrin. 2010. Performance of recommender algorithms on top-n recommendation tasks. In *Proceedings of the fourth ACM conference on Recommender systems.* ACM, 39–46.

[3] Michael D Ekstrand, John T Riedl, and Joseph A Konstan. 2011. Collaborative filtering recommender systems. *Foundations and Trends in Human-Computer Interaction* 4, 2 (2011), 81–173.

[4] A Felfernig, S Gordea, D Jannach, E Teppan, and M Zanker. 2007. A short survey of recommendation technologies in travel and tourism. *OEGAI journal* 25, 7 (2007), 17–22.

[5] Jonathan L Herlocker, Joseph A Konstan, and John Riedl. 2000. Explaining collaborative filtering recommendations. In *Proceedings of the 2000 ACM conference on Computer supported cooperative work.* ACM, 241–250.

[6] Antonio Hernando, JesúS Bobadilla, Fernando Ortega, and Abraham GutiéRrez. 2013. Trees for explaining recommendations made through collaborative filtering. *Information Sciences* 239 (2013), 1–17.

[7] Marius Kaminskas and Francesco Ricci. 2012. Contextual music information retrieval and recommendation: State of the art and challenges. *Computer Science Review* 6, 2 (2012), 89–119.

[8] Qi Liu, Enhong Chen, Hui Xiong, Yong Ge, Zhongmou Li, and Xiang Wu. 2014. A cocktail approach for travel package recommendation. *IEEE Transactions on Knowledge and Data Engineering* 26, 2 (2014), 278–293.

[9] Pasquale Lops, Marco Gemmis, and Giovanni Semeraro. 2011. Recommender Systems Handbook. Content-based Recommender Systems: State of the Art and Trends (2011), 73–105. http://dx.doi.org/10.1007/978-0-387-85820-3_3

[10] X. Ning, C. Desrosiers, and G. Karypis. 2015. *A comprehensive survey of neighborhood-based recommendation methods.* 37–76.

[11] Xiaoyuan Su and Taghi M Khoshgoftaar. 2009. A survey of collaborative filtering techniques. *Advances in artificial intelligence* 2009 (2009), 4.

[12] Panagiotis Symeonidis, Alexandros Nanopoulos, and Yannis Manolopoulos. 2008. Justified recommendations based on content and rating data. In *WebKDD Workshop on Web Mining and Web Usage Analysis.*

[13] Nava Tintarev and Judith Masthoff. 2012. Evaluating the effectiveness of explanations for recommender systems. *User Modeling and User-Adapted Interaction* 22, 4-5 (2012), 399–439.

[14] Nava Tintarev and Judith Masthoff. 2015. Explaining recommendations: Design and evaluation. In *Recommender Systems Handbook.* Springer, 353–382.

[15] Jesse Vig, Shilad Sen, and John Riedl. 2009. Tagsplanations: explaining recommendations using tags. In *Proceedings of the 14th international conference on Intelligent user interfaces.* ACM, 47–56.

[16] Beidou Wang, Martin Ester, Jiajun Bu, and Deng Cai. 2014. Who Also Likes It? Generating the Most Persuasive Social Explanations in Recommender Systems.. In *AAAI.* 173–179.

[17] A. T. Wibowo, A. Siddharthan, H. Anderson, A. Robinson, Nirwan Sharma, H. Bostock, A. Salisbury, R. Comont, and R. V. D. Wal. 2017. Bumblebee Friendly Planting Recommendations with Citizen Science Data. In *Proceedings of the RecSys 2017 Workshop on Recommender Systems for Citizens co-located with 11th ACM Conference on Recommender Systems (RecSys 2017), Como, Italy, August 31, 2017.*

[18] A. T. Wibowo, A. Siddharthan, C. Lin, and J. Masthoff. 2017. Matrix Factorization for Package Recommendations. In *Proceedings of the RecSys 2017 Workshop on Recommendation in Complex Scenarios co-located with 11th ACM Conference on Recommender Systems (RecSys 2017), Como, Italy, August 31, 2017.* 23–28. http://ceur-ws.org/Vol-1892/paper5.pdf

[19] Agung Toto Wibowo, Advaith Siddharthan, Judith Masthoff, and Chenghua Lin. 2018. Incorporating Constraints into Matrix Factorization for Clothes Package Recommendation. In *Proceedings of the 26th Conference on User Modeling, Adaptation and Personalization.* ACM.