

Selecting News to Suit a Group of Criteria: An Exploration

Judith Masthoff

University of Brighton
Brighton, UK
Judith.Masthoff@brighton.ac.uk

Abstract. In (Masthoff, 2004), we have investigated techniques for combining individual user models to make recommendations to a group. In Masthoff (2003), we have shown how these techniques might also be applicable when adapting to an individual: for solving the cold start problem and for combining ratings on multiple criteria. In this paper, we look at combining multiple criteria in more detail. We present an exploratory experiment into how people combine criteria. A main issue is potential inequality of the criteria's importance. We show how the group modelling strategies can be adapted to deal with this inequality.

Introduction

One problem faced by personalized TV is that viewers often watch television in groups. Hence, previously we have studied adaptation to groups of users, in particular strategies for combining individual user models in order to determine group recommendations (Masthoff, 2004). For instance, suppose the TV knows who is watching, and it knows the ratings of each individual for a set of news items, then there are various strategies for deciding which news items to show to the group. In (Masthoff, 2003), we have claimed that the usefulness of these strategies is not restricted to adaptation to groups of people, but that these strategies could also be useful when adapting to an *individual*. We have shown how group modelling could contribute to solving the cold start problem: the problem that the TV does not know enough about the user initially to make good recommendations. If there is a set of known other users, or a set of stereotypes, and the system does not yet know which of these the user resembles, then one way to ensure enjoyment is to present items that the other users (or stereotypes) would enjoy as a *group*. We have shown how this could work on some real data. We have also discussed how the group modelling strategies might be applicable when combining ratings on multiple criteria. Multiple factors can contribute to a recommender's prediction of the user's opinion of an item. For instance, Pazzani (1999) *adds* the rankings produced by content-based filtering, collaborative filtering, and demographic filtering. Nguyen and Haddawy (1998) use a *weighted addition* of attributes describing a movie: director, casting, genre, star rating and running time. Though stressing the importance of aggregating ratings, neither

discusses why they have chosen their particular aggregation function. In Masthoff (2003), we have claimed that the same strategies could be used as when combining individual ratings for a group model, however, without evidence to back this up. In this paper, we will explore what really happens when ratings on multiple criteria are to be combined. Our example domain will be personalized news, which is one of the more popular areas for research into Personalized TV (e.g., Maybury et al., 2004; Dimitrova et al., 2004).

Summary of Group Modelling Strategies

In this section, we will summarize the strategies subjects seemed to use in the group modelling experiment in Masthoff (2004). For more details and an overview of additional strategies see Masthoff (2004). We had asked subjects to recommend a sequence of items to a group, based on item ratings of the individual group members. In the following, the so-called ‘group list’ is the sequence in which items would be chosen by a particular group modelling strategy. Sometimes, two items score the same, like E and F in the average strategy. This is indicated in the group list by placing them between brackets. This means that either E is followed by F, or F is followed by E.

Three strategies mimicked the behaviour of many of our subjects:

1. *Average strategy (also called Additive Utilitarianism)*. Ratings are added, and the larger the sum the earlier the alternative appears in the sequence (results are the same as when averaging ratings, hence its name). This strategy (often in a weighted form, where weights are attached to individual ratings) is also used in multi-agent systems (Hogg & Jennings, 1999), Collaborative filtering, and in the INTRIGUE system with recommends attractions to visit to groups of tourists (Ardissono et al, 2002). A disadvantage is that starvation can occur: if a person’s opinions differ from those of most people in the group, they might never get anything they like.

	A	B	C	D	E	F	G	H	I	J
John	10	4	3	6	10	9	6	8	10	8
Adam	1	9	8	9	7	9	6	9	3	8
Mary	10	5	2	7	9	8	5	6	7	6
Group	21	18	13	22	26	26	17	23	20	22

Group List:
(E,F)H(D,J)AIBGC

2. *Least Misery Strategy*. The minimum of the individual ratings is taken, and the larger the minimum the earlier the alternative appears in the sequence. The idea behind this strategy is that a group is as happy as its least happy member. The POLYLENS movie recommender system uses this strategy, assuming groups of people going to watch a movie together tend to be small and a small group to be as happy as its least happy member (O’Conner, Cosley, Konstan & Riedl, 2001). A disadvantage is that a minority opinion can dictate the group: if everybody really wants to see something, but one person does not like it, then it will never be seen.

	A	B	C	D	E	F	G	H	I	J
John	10	4	3	6	10	9	6	8	10	8
Adam	1	9	8	9	7	9	6	9	3	8
Mary	10	5	2	7	9	8	5	6	7	6
Group	1	4	2	6	7	8	5	6	3	6

Group List:
FE(H,J,D)GBICA

3. *Average Without Misery Strategy*. Ratings are added, but without items that score below a certain threshold (say 4) for individuals. MUSICFX (McCarty & Anagnost, 1998) uses a slightly more complex version of this strategy, when selecting music for people working out in a fitness center.

	A	B	C	D	E	F	G	H	I	J
John	10	4	3	6	10	9	6	8	10	8
Adam	1	9	8	9	7	9	6	9	3	8
Mary	10	5	2	7	9	8	5	6	7	6
Group	-	18	-	22	26	26	17	23	-	22

Group List
threshold 4:
(E,F)H(D,J)BG
threshold 3):
(E,F)H(D,J)IB

In Masthoff (2004), we found that subjects cared about avoiding misery and about fairness. Many subjects' behaviour reflected that of the Least Misery or Average Without Misery strategies. Even subjects whose behaviour completely deviated from these strategies, avoided

Experiment: How People Combine Multiple Criteria

We wanted to investigate whether the way subjects combine ratings from multiple criteria reflects that of subjects combining ratings of individual group members. Hence, the experimental design used was identical to that in Masthoff (2004), and the same ratings were used. The only difference was that this time we used criteria rather than group members.

Method

Subjects were given the individual ratings of three criteria for a set of news items. The criteria used were:

- *Importance*. How important the news is in general. For instance, a major earthquake is more important than a single house collapsing.
Each item is rated from 1 –really unimportant- to 10-really important.
- *Relevance of Location*. How relevant the location of the news is to you. For instance, for a Dutch person living in Brighton, the relevance of location of news from the Netherlands and Brighton is higher than that of news from Italy.
Each item is rated from 1 –really irrelevant location- to 10-really relevant location.
- *Recency*. How recent the news is. For instance, something that happened in the last hour is more recent than something that happened yesterday.
Each item is rated from 1 –really old news- to 10-really recent news.

Bell (1991) provides an overview of criteria used by editors for news selection. He mentions more than 15 criteria. The above are inspired by those. For instance, our Recency criterion is a combination of his Recency and Freshness criteria. Our Relevance of Location criterion is a combination of his Proximity and Relevance criteria. Note that we did not explicitly set out to create criteria of different importance, though the experimental results seem to show such a difference.

In seven questions, subjects were asked which item they would watch, given that they only had time to see respectively 1, 2, 3, 4, 5, 6, or 7 items, and why they made that selection (see Appendix A for exact task wording). The individual ratings used were the same as those that had been used in the Experiments in Masthoff (2004), with Importance having the ratings of John, Relevance of Location having the ratings of Adam, and Recency having those of Mary. See Table 1 for the ratings used.

Table 1. Ratings used in the experiment. A to J are news items.

	A	B	C	D	E	F	G	H	I	J
Importance	10	4	3	6	10	9	6	8	10	8
Relevance of Location	1	9	8	9	7	9	6	9	3	8
Recency	10	5	2	7	9	8	5	6	7	6

The ratings had been chosen primarily to enable differentiating between the strategies we expected subjects to use. In addition, it ensured that Importance and Recency had quite similar ratings, while the Relevance of Location's ratings were frequently the opposite of the ratings of the other two. We also ensured that for one news item, namely item A, Importance and Recency had maximal positive ratings (10), while Relevance of Location had a maximal negative rating (1). The latter would give a good idea of the importance subjects assigned to avoiding misery.

Subjects

Thirty six subjects participated in the experiment (92% male, 8% female, average age 23.8, standard deviation 4.6). All were final-year undergraduate students in computing attending a lecture of the Adaptive Interactive Systems module. The experiment took place in a lecture room. Students participated in the experiment voluntarily (in addition to the numbers mentioned above, 4 students chose not to participate).

Results and Discussion

Subjects do not seem to answer the questions independently: they responded which new item should be added to the sequence they had already chosen for the previous question. This made it possible to present the results in the way we have done in Table 2, only showing the new item selected for each question. We have tried to keep the table as simple (and uncrowded) as possible: if a cell does not have an item name in it, then the first name above it applies. For instance, s1 replied F to the first question. Subjects have been ordered to make the table as easy to view as possible.

Table 2. Results. See below for meaning of shading and border lines.

	1	2	3	4	5	6	7	Summary	
s5	F	E	A	I	J	D	H	None of the discussed strategies.	
s7					H		B		
s33						J	D		
s22									
s1				H	I		?		
s29					J	D	B		
s20				D	H	I	J		
s13			H	J	D	A	I		Average strategy throughout.
s28									
s25				D	B		J		Average strategy or Least Misery strategy for first four items
s3		A	E	I	H	J	D	None of the discussed strategies.	
s19		J		A	D	H	I		
s4	E	F	A	I	H	J	D		
s23					J	H			
s24							B		
s26				J	H	I	D		
s16				H	I	J			
s6			H	J	D	B	I		Average strategy for the first five items.
s8							G		Average without Misery throughout.
s10						A	I		Average strategy throughout.
s12					I		D	Average strategy for the first four items.	
s2									
s30				D	A	I	J	None of the discussed strategies.	
s15			J	H	D	A	I		
s27			I	A	J	H	G		
s11		A	F	I	H	J			
s21				H	D	B	J		
s14	A	E	I	F	H	J	D		
s31			F	I			B		
s32						D	J		
s35					J	H	D		
s36			D	F	I		J		
s17		F	E	I	H	J	D		
s18				H	J	D	I		
s34		I		J	F	H	?		

The table includes information about how well the subjects' replies fit the strategies discussed above:

- Bold borderlines indicate replies that are in correspondence with the Average Strategy. So, for instance, all replies of s13 were the same as those by the Average Strategy. The first two replies by s4 were the same as those by the Average Strategy, but s4's later replies differed.
- Gray cell shading indicates replies that are in correspondence with the Least Misery Strategy. So, for instance, the first two replies by s5 were the same as those by the Least Misery Strategy, but s5's later replies differed.
- Bold dotted borderlines indicate replies that are in correspondence with the Average Without Misery Strategy. So, for instance, all replies of s8 were the same as those by the Average Without Misery Strategy.

Note that strategies can have overlapping starts of their group lists. For instance, both the Average Strategy and the Least Misery Strategy allow a start of FEHDJ. This means that cells can have both grey cell shading and a bold borderline. So, for instance, s13's replies followed the Least Misery Strategy for the first five items, and were in correspondence with the Average Strategy for all seven items. Also, we have only used the Bold dotted borderlines, when the Average Without Misery Strategy starts deviating from the Average Strategy. The results of one subject, s9, are excluded from the discussion. He misinterpreted the ratings, believing 1 to mean good and 10 bad, calling item C "the most important news".

Comparing the results in Table 2 with those in (Masthoff, 2004), the most striking difference is that the results of subjects in (Masthoff, 2004) tended to reflect those of strategies, particularly the Average Strategy and Least Misery Strategy. In contrast, hardly any subjects in this experiment seem to have followed one of the discussed strategies. Particularly, we find hardly any evidence of avoiding misery, which was found to be an important factor for subjects when making group recommendations. This may be due to subjects attaching different weighting to the criteria. Many subjects talk about the Importance and Recency of items when explaining their choice, in contrast to hardly any mention of the Relevance of Location. One subject (s34) even states explicitly that he "is not that interested in local news".

As can be seen in Figure 1, item A and I are clearly introduced earlier in this experiment than in the experiment on adaptation to a group of people reported in (Masthoff, 2004). This indicates that subjects indeed use different weightings, and attach less weight to the "Relevance of Location" criterion than to the other criteria. As many as nine subjects introduce J (ratings 8-8-6) before H (ratings 8-9-6). Again, this seems to show a disregard for criterion 2 "Relevance of Location".

Subjects sometimes seem to interpret ratings, mapping them onto real life news items. For instance, three subjects (S14, S18, and S33) described item A as "September the 11th", a news item of great importance and very recent (at the time of broadcast), but related to (according to them) a less relevant location. Some subjects seem to equate the "Relevance of Location" criterion with whether the news is local or not. In the UK, there are local news broadcasts after the main news, with news items like "elderly lady mugged in Brighton". This might have influenced subjects' opinion of this criterion.

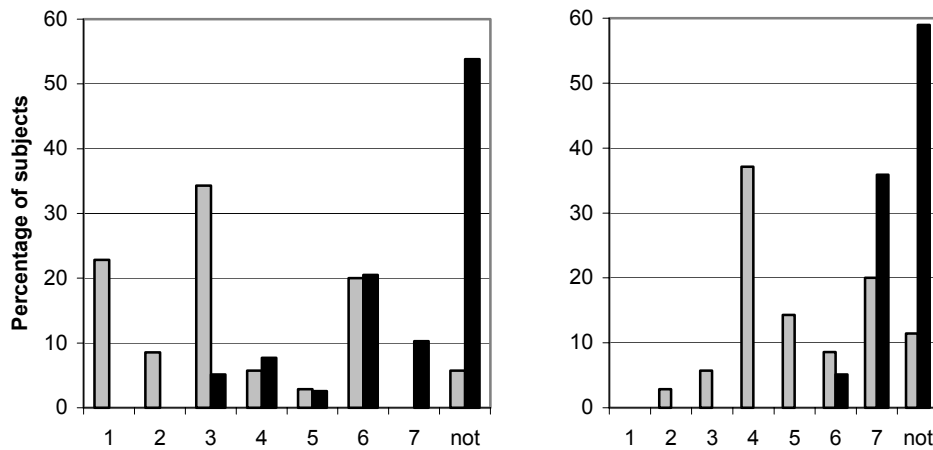


Fig. 1. Introduction of items A (left) and I (right) in this experiment (grey) and in the old experiment (adaptation to a group of people) (black).

Adapting the Strategies to Deal with Inequality

In Masthoff (2004), we noted that individuals in a group can be of varied importance, and discussed how weightings could be used in an Average Strategy, or even how only the ratings of the most important individual could be used (we called this the Most Respected Person strategy). In our experiments in (Masthoff, 2004), however, subjects regarded all individuals as equally important, which is not surprising given that they were only provided with their names. (We tried to influence subjects' attribution of importance in one experiment by giving John, Adam and Mary varying ages, but without success). Actually, it seems quite sensible and morally correct to treat everybody equally when a system is adapting to a group of people. Of course, there may be some exceptions, for instance when the group contains adults as well as children, or when it is somebody's birthday. But in general, equality seems a good choice. In contrast, when a system is adapting to a group of criteria, rather than people, there is no particular reason for assuming all criteria are as important. It is even quite likely that not all criteria are equally important for a particular person. The result of the experiment discussed above do indeed show such an inequality between criteria. So, how can we adapt the strategies to deal with this, and to better explain the behaviour of our subjects? There are several ways in which this can be done:

1. *Average Of Important Criteria and Least Misery For Important Criteria*

The ratings of unimportant criteria are ignored completely. For instance, assume criterion Location is regarded unimportant, then its ratings are ignored. The result of the Average Of Important Criteria strategy becomes:

	A	B	C	D	E	F	G	H	I	J
Importance	10	4	3	6	10	9	6	8	10	8
Recency	10	5	2	7	9	8	5	6	7	6
Group	20	9	5	13	19	17	11	14	17	14

Group List:
AE(F,I)(H,J)DGBC

And the result of the Least Misery For Important Criteria strategy becomes:

	A	B	C	D	E	F	G	H	I	J
Importance	10	4	3	6	10	9	6	8	10	8
Recency	10	5	2	7	9	8	5	6	7	6
Group	10	4	2	6	9	8	5	6	7	6

Group List:
AEFI(H,I,D)GBC

2. Average With Weights.

The ratings of unimportant criteria are given less weight. The weight of a criterion is multiplied with its ratings to produce new ratings. For instance, suppose criteria Importance and Recency were three times as important as criterion Location. The result of the Average With Weights Strategy becomes:

	A	B	C	D	E	F	G	H	I	J
Importance	30	12	9	18	30	27	18	24	30	24
Location	1	9	8	9	7	9	6	9	3	8
Recency	30	15	6	21	27	24	15	18	21	18
Group	61	36	23	48	64	60	39	51	54	50

Group List
weight 3-1-3:
EAFIHJDGBC
weight 2-1-2:
EFA(H,I)JDGBC

Note that a weight of 0 results in ignoring the ratings completely, as above.

3. Average Without Misery For Some.

Misery is avoided for important criteria but not for unimportant ones. Assume criterion Location is again regarded as unimportant. The result of the Average Without Misery For Some strategy with threshold 6 becomes:

	A	B	C	D	E	F	G	H	I	J
Importance	10	4	3	6	10	9	6	8	10	8
Location	1	9	8	9	7	9	6	9	3	8
Recency	10	5	2	7	9	8	5	6	7	6
Group	21			22	26	26		23	20	22

Group List
threshold 6:
(EF)H(J,D)AI
threshold 7:
(EF)AI

Table 3 shows how the subjects' replies fit with the strategies discussed above:

- Bold borderlines indicate replies that are in correspondence with the Average strategy (thin line) or the Average With Weights strategy for weights 2-1-2 or 3-1-3 (fat line).
- Bold dotted borderlines indicate replies that are in correspondence with the Average Without Misery strategy (small dots), or the Average Without Misery For Some strategy (large dots) with criterion Location as unimportant.
- Shading indicates replies that are in correspondence with the Least Misery strategy (light grey), or the Least Misery for Important Criteria strategy (dark grey) with criterion Location as unimportant. Subject s14 explicitly indicated that he was using this strategy, stating that he always went for the item with the highest Importance, disregarding the other criteria.

Some replies correspond to multiple strategies. E.g., subject s4's response corresponds both to Average with Weights and Average Without Misery For Some. This has been indicated in the summary column.

Table 3. Results mapped onto the strategies with inequality. See above for meaning of shading and borderlines.

	1	2	3	4	5	6	7	Summary
s5	F	E	A	I	J	D	H	Average Without Misery For Some threshold 7 followed by threshold 6
s7					H		B	
s33						J	D	
s22								
s1				H	I		?	Average Without Misery For Some for the first three items. Threshold 7.
s29					J	D	B	
s20				D	H	I	J	
s13			H	J	D	A	I	Average strategy throughout.
s28								
s25				D	B		J	Average strategy or Least Misery strategy for first four items
s3		A	E	I	H	J	D	None of the discussed strategies.
s19		J		A	D	H	I	
s4	E	F	A	I	H	J	D	Average with Weights, 2-1-2. Or, Average Without Misery For Some, t7-t6.
s23					J	H		For first four: Average with Weights 2-1-2. Or Average Without Misery For Some, t7.
s24							B	
s26				J	H	I	D	Average with Weights, 2-1-2, throughout.
s16				H	I	J		
s6			H	J	D	B	I	
s8							G	Average Without Misery throughout.
s10						A	I	Average strategy throughout.
s12					I		D	Average strategy for the first four items.
s2								
s30				D	A	I	J	
s15			J	H	D	A	I	None of the discussed strategies.
s27			I	A	J	H	G	
s11		A	F	I	H	J		Average with Weights, 3-1-3, for first six.
s21				H	D	B	J	
s14	A	E	I	F	H	J	D	Least Misery (or Average) for Important Criteria: Importance only.
s31			F	I			B	Least Misery for Important Criteria
s32						D	J	
s35					J	H	D	None of the discussed strategies.
s36			D	F	I		J	
s17		F	E	I	H	J	D	
s18				H	J	D	I	
s34		I		J	F	H	?	

Comparing Table 2 with Table 3, we conclude that our adaptation of the strategies to cope with inequality has had some success in explaining subject behaviour. There also seems to be some evidence now that subjects still do care about avoiding misery. Subjects just do not seem to care about the criterion Relevance of Location, and therefore do not care about avoiding misery for that criterion.

Conclusions

The problem of combining rating for multiple criteria bears a similarity with that of combining rating for multiple people. We have performed an exploratory experiment to investigate to what extent results from our group modelling work can be used when combining multiple criteria, for instance for personalized news. The results show that a main issue when combining criteria is dealing with a likely inequality of the criteria's importance. We have shown some ways in which the strategies can be modified to cope with this. This is only the first step in the research. The next step would be to show subjects sequences generated by our algorithms and ask them to rate the user's predicted satisfaction.

References

- Ardissono, L., Goy, A., Petrone, G., Segnan, M., and Torasso, P. (2002). Tailoring the recommendation of tourist information to heterogeneous user groups. In S. Reich, M. Tzagarakis, and P. De Bra (eds.), *Hypermedia: Openness, structural awareness, and adaptivity*, International Workshops OHS-7, SC-3, and AH-3, 2001. Lecture Notes in Computer Science 2266, Berlin: Springer Verlag, pp. 280-295.
- Bell, A. (1991). *The language of news media*. Oxford, UK: Basil Blackwell.
- O' Conner, M., Cosley, D., Konstan, J.A., and Riedl, J. (2001). PolyLens: A recommender system for groups of users. In: *Proceedings of ECSCW 2001*, Bonn, Germany, pp. 199-218. As accessed on <http://www.cs.umn.edu/Research/GroupLens/poly-camera-final.pdf>.
- Dimitrova, N., Zimmerman, J., Janevski, A., Agnihotri, L., Haas, N., Li, D., Bolle, R., Velipasalar, S., McGee, T., and Nikolovska, L. (2004). Media augmentation and personalization through multimedia processing and information extraction. In L.Ardissono, A. Kobsa and M. Maybury (eds.), *Personalized digital television: Targeting programs to individual viewers*. Dordrecht, NL: Kluwer, pp203-233.
- Hogg, L., and Jennings, N.R. (1999). Variable sociability in agent-based decision making. Sixth International Workshop on Agent Theories, Architectures and Languages, Orlando, FL, USA, pp. 276-289.
- Masthoff, J. (2004). Group modeling: Selecting a sequence of television items to suit a group of viewers. *User Modeling and User Adapted Interaction*, 14, pp37-85.
[Also published in L.Ardissono, A. Kobsa and M.Maybury (eds.), *Personalized digital television: Targeting programs to individual viewers*. Dordrecht, NL: Kluwer.]
- Masthoff, J. (2003). Modeling the multiple people that are me. In: P. Brusilovsky, A. Corbett, and F. de Rosi (eds.) *Proceedings of the 2003 User Modeling Conference*, Johnstown, PA, Berlin: Springer Verlag, pp.258-262.
- Maybury, M., Greiff, W., Boykin, S., Ponte, J., McHenry, C, and Ferro, L. (2004). Personalcasting: Tailored broadcast news. *User Modeling and User Adapted Interaction*, 14,

pp119-144. [Also published in L.Ardissono, A. Kobsa and M.Maybury (eds.), Personalized digital television: Targeting programs to individual viewers. Dordrecht, NL: Kluwer]

McCarthy, J., and Anagnost, T. (1998). MusicFX: An arbiter of group preferences for computer supported collaborative workouts. *ACM 1998 Conference on CSCW*, Seattle, WA, pp. 363-372.

Nguyen, H., and Haddawy, P. (1998). The decision-theoretic video advisor. *Proceedings of AAAI Workshop on Recommender Systems*, Madison, WI. pp. 77-80.

Pazzani, M.J. (1999). A framework for collaborative, content-based and demographic filtering. *Artificial Intelligence Review*, 13, pp. 393-408.

Appendix A: Task Wording

You are going to watch the news. For each news item, we know

- How important the news is in general. For instance, a major earthquake is more important than a single house collapsing.
Each item is rated from 1 –really unimportant- to 10-really important.
- How relevant the location of the news is to you. For instance, for a Dutch person living in Brighton, the relevance of location of news from the Netherlands and Brighton is higher than that of news from Italy.
Each item is rated from 1 –really irrelevant location- to 10-really relevant location.
- How recent the news is. For instance, something that happened in the last hour is more recent than something that happened yesterday.
Each item is rated from 1 –really old news- to 10-really recent news.

News Item	Importance	Relevance of Location	Recency
A	10	1	10
B	4	9	5
C	3	8	2
D	6	9	7
E	10	7	9
F	9	9	8
G	6	6	5
H	8	9	6
I	10	3	7
J	8	8	6

1. You have only time to watch one item. Which item would you watch? Why?
2. You have only time to watch two items. Which items would you watch? Why?
3. You have only time to watch three items. Which items would you watch? Why?
4. You have only time to watch four items. Which items would you watch? Why?
5. You have only time to watch five items. Which items would you watch? Why?
6. You have only time to watch six items. Which items would you watch? Why?
7. You have only time to watch seven items. Which items would you watch? Why?