Initial Perspectives From Preferences Expressed Through Comparisons

Nicolas Jones, Armelle Brun, and Anne Boyer

LORIA - Nancy Université, BP 239, 54506 Vandœuvre-lès-Nancy, France
{nicolas.jones, armelle.brun, anne.boyer}@loria.fr,

Abstract. Rating-scales have become a popular modality for expressing our preferences, but they present several drawbacks. We have recently proposed a new modality: comparing items ("I prefer A to B"). After initial user-studies with encouraging results, we here share some initial perspectives. In particular we examine three issues illustrated with graphs of user's preferences. We discuss the adaptability of comparisons, their algorithmic complexity and incoherences introduced by transitivity.

Keywords: preference expression, comparisons, ratings

1 Introduction: Context for Using Comparisons

Multi-point rating scales have become a popular preference expression tool for personalization and recommender systems. They unfortunately present several drawbacks: users' ratings are inconsistent through time [1]; the issue of the optimal number of points in rating scales is still unresolved [7]; the granularity of the scale often offers limited precision, which likely to frustrate users [4]; the values and descriptive labels associated with the scale points can influence users's ratings [2], etc.

We recently proposed a new modality whereby users compare items two-by-two "I prefer A to B" ($A \rightarrow B$), as one often does in everyday life. To evaluate this alternative, we ran several user studies with over 200 participants [6, 5], where these got to rate or compare films and television series. Three important findings were obtained. First, preferences expressed with comparisons are coherent with those from ratings, although some differences appear. When users say that they prefer a movie A to B with a comparison, the rating scores reflect this preference in 92.5% of cases. However, when both are equal in comparisons ($A \leftrightarrow B$), ratings are only equal 42.7% of times. Second, participants preferred comparisons, and were favorably predisposed to using them instead of ratings. Users found comparisons easier to use, requiring less effort, but at the same time found the ratings to give more control. Third, over a fifteen day break, comparisons are 20% more stable through time than ratings.

Despite these favorable findings, certain issues about comparisons need to be raised. In this paper, we discuss three high-level considerations relating to preferences expressed through comparisons. In order to model users' preferences, we used *preference relations* to create ranked graphs of users' comparisons [3]. Several graphs are presented in this paper.

2 Discussion

2.1 Adaptability of Comparisons

One issue with rating scales is that, within a system, they only provide a fixed range of possible answers. Depending on the situation, users might wish to have more or fewer points on the scale. We expected that comparisons would be more flexible, since users can explicitly say that two items should be equivalent or if a difference is perceived. This adaptability was confirmed by our results. These show that some users need only three levels of rankings in their preference relation, whereas others go up to nine levels; the median score is 5 ranks. Examples with respectively three and seven levels are shown in Figure 3 and Figure 1. One should keep in mind that these results were only obtained over a three minute session, and that more detailed relations are likely to appear over time.

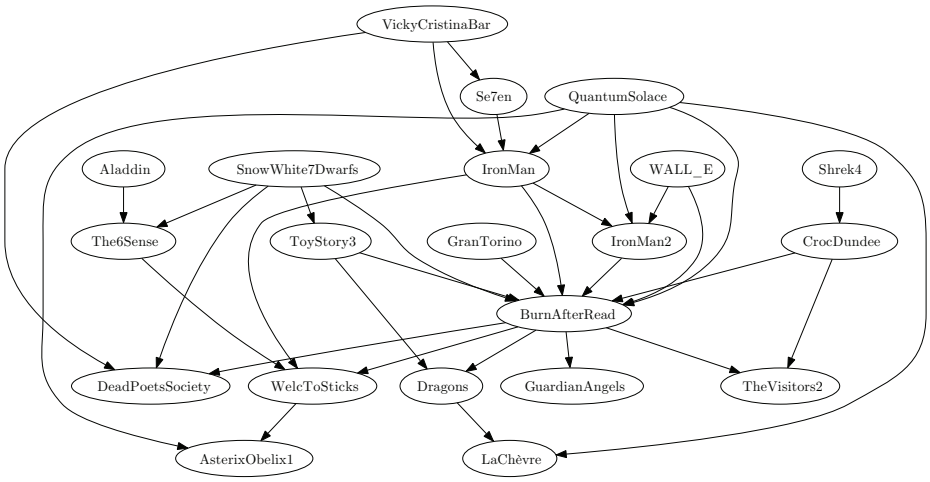

Fig. 1. Example of a users' preferences with seven ranks.

More importantly, we believe that the adaptability of comparisons means that they are more precise at modeling users' preferences than ratings. Let us imagine the case of users who only need three levels to rank items. If they are given a five-point rating scale, it is highly likely that they will give neighboring scores to items that they actually perceive as being indifferent. In doing so, they will introduce noise into their data. This phenomena is known in literature and was reported by several users during post-study discussions. In contrast, comparisons seem to allow users to stay on three levels, and therefore they model users' preferences more precisely. If we consider the opposite case, of users who rely on more than five levels, they are often obliged to give the same rating to two films which they actually appreciate differently. Here again the comparisons seem to solve

this issue, with some users relying on up to nine ranks of preferences, therefore modeling users' preferences more precisely. For these reasons we are convinced that comparisons will increase the quality of the user information recorded.

2.2 Complexity of Comparisons

A second issue about comparisons is their algorithmic complexity. In order to obtain a complete comparison graph, where each item has been compared to all others once, $n * (n-1)/2$ comparisons are needed, where n is the number of items. Whilst this theoretical number is very high and thus an obvious drawback, encouraging signs show that not all comparisons are useful. In a full preference relation, collected for instance over a three-minute session, some items may have a high uncertainty when determining their ranking. To illustrate this issue, we portray two examples in Figure 2: we consider two similar items M and N that have the same relation towards E and F. By choice, we positioned the first item M at rank 3 with a high uncertainty, as it could just as well be on either of the two ranks above it (M' or M"). However, the promising observation is that with just one additional comparison the doubt around N can be resolved. If we compare N to D we will not learn anything new, whilst a comparison with B (labelled j) will allow to position N with confidence.

For this reason, we believe that by using an adequate strategy for selecting pairs to be compared, we can easily limit the complexity of comparisons. We are confident that a compromise between completeness of the preference relation, and asking users to compare items a minimal number of times, can be reached. This should especially be true, as our results show that users find it easier, and requiring less effort, to make comparisons rather than ratings. A possible strategy for selecting which comparisons should be made would be to use dichotomy among the ranks with uncertainty.

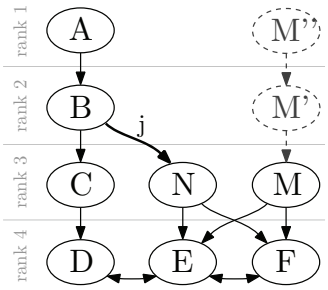

Fig. 2. Uncertainty of rankings illustrated.

2.3 Incoherences in Comparisons

A third challenge about using comparisons is incoherences induced by transitivity. In Section 2.2 we supposed that transitivity could be used to deduce unknown preferences. In order to reflect on how realistic this claim is, we here discuss the incoherences that users created in their three minute sessions of comparisons.

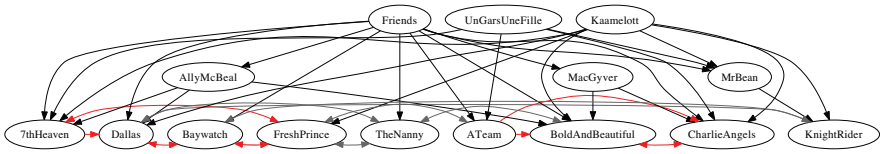

Fig. 3. Example of a users' preferences with three ranks, and incoherences.

We consider that a user has introduced an incoherence when a cycle in the preference relation can be detected among several comparison couples among which at least one is not a strict equivalence. (for example: A→B→C→A). In one of our experiments, less than half of the users presented cycles with incoherences, the majority of them only having one or two incoherences throughout their whole graph. This shows that within three-minute sessions, users are mainly coherent.

When users' graphs present incoherences, this can seem problematic. However, having run these preliminary studies, and considering users' feedback, we are under the impression that these incoherences should not be perceived only as a weakness, but rather as an opportunity to improve and refine the user-model. Currently two strategies appear to us as being easily implementable.

A first strategy would be to weight edges. We observed that most of the incoherences appeared when users indicated one or more equivalence relations inside a cycle. Figure 3 shows two examples, highlighted in red. When multiple items are judged as being equivalent, it seems that there is an increased uncertainty in the use of transitivity across the current rank. Said otherwise: users are able to make equivalence judgements between two items at a time, but a sequence of equivalences does not guarantee that all items are on the same global rank. It is conceivable that multiple equivalencies could spread across two ranks. We believe that such incoherences across equivalences could be detected and that the weights of the graph, or the ranks of the items, could be tweaked to create a finer user-model.

Second, another possible strategy would be to simply remove conflicting edges in the system. In the field of database, a recent work by Pyratos *et. al* used the same strategy, leading to the non-inclusion of new edges creating an incoherence [8]. A variant of this approach might be to keep the last added comparison, and to try and remove the oldest comparison(s) which created this incoherence. In doing so, the system would give priority to the users current feedback, a common tactic in the field of user modeling.

3 Conclusions

The points discussed in this paper show that comparisons are a highly promising modality for expressing preferences. We use preference relations to model users' preferences as ranked graphs. Using these graphs, we first observe that the preferences expressed with comparisons are very adaptive to users' needed level of detail. Second, although the number of comparisons required to build a reliable preference relation is high, we show that not all comparisons are useful and that an adequate strategy may dramatically reduce the number of comparisons required. Third, we point out that users don't express many incoherences through a three-minute session of comparisons, and that those observed often appear in equivalence relations. We argue that incoherences could present an additional value, a mechanism for weighting users' graphs. These findings conduct us to the following statement: when using rating scales, the number of answers is limited, which reduces the precision of the preferences expressed, but facilitates their automatic processing. At the opposite, when using comparisons, a finer precision of preferences is obtained, but these are more complex to process. We believe that if we can make comparisons an even easier interaction-task, and deduce comparisons from users' traces, comparisons will become a much more valuable preference-expression mechanism than the current rating-scales. These elements constitute our future work.

References

1. Amatriain, X., Pujol, J.M., Oliver, N.: I like it... i like it not: Evaluating user ratings noise in recommender systems. In: Proceedings of the User Modeling, Adaptation, and Personalization (UMAP'09) (2009)
2. Amoo, T., Friedman, H.: Do numeric values influence subjects' responses to rating scales? Journal of International Marketing and Marketing Research 26, 41–46 (2001)
3. Brun, A., Hamad, A., Buffet, O., Boyer, A.: Towards preference relations in recommender systems. In: Workshop on Preference Learning, European Conference on Machine Learning and Principle and Practice of Knowledge Discovery in Databases (ECML-PKDD 2010) (2010)
4. Cena, F., Vernero, F., Gena, C.: Towards a customization of rating scales in adaptive systems. In: Proceedings of the User Modeling, Adaptation, and Personalization (UMAP'10). pp. 369–374 (2010)
5. Jones, N., Brun, A., Boyer, A.: Comparisons Instead of Ratings: Towards More Stable Preferences. Tech. rep., LORIA, Nancy Université (2011)
6. Jones, N., Brun, A., Boyer, A.: Ratings, What Else? Tech. rep., LORIA, Nancy Université (2011)
7. Preston, C., Colman, A.: Optimal number of response categories in rating scales: reliability, validity, discriminating power, and respondent preferences. Acta Psychologica 104(1), 1–15 (2000)
8. Spyratos, N., Kotzinos, D.: Communicating through preferences. In: PETRA (2010)