An Exploratory Work in Using Comparisons Instead of Ratings

Nicolas Jones[1], Armelle Brun[1], Anne Boyer[1], Ahmad Hamad[2]

[1]LORIA - Nancy Université, KIWI Group
[2] Sailendra SAS
BP 239, 54506 Vandœuvre-lès-Nancy, France
{nicolas.jones,armelle.brun,anne.boyer,ahmad.hamad}@loria.fr

Abstract. With the evolution of the Web, users are now encouraged to express their preferences on items. These are often conveyed through a rating value on a multi-point rating scale (for example from one to five). Ratings have however several known drawbacks, such as imprecision and inconsistency. We propose a new modality to express preferences: comparing items ("I prefer x to y"). In this initial work, we conduct two user-studies to understand the possible relevance of comparisons. This work shows that users are favorably predisposed to adopt this new modality. Moreover, it shows that preferences expressed as ratings are coherent with preferences expressed trough comparisons, and to some extent equivalent. As a proof of concept, a recommender is implemented using comparison data, where we show encouraging results when confronted to a classical rating-based recommender. As a consequence, asking users to express their preferences through comparisons, in place of ratings, is a promising new modality for preference-expression.

1 Introduction

With the emergence of the Web 2.0, users are encouraged to express their preferences about events, websites, products, etc. These preferences are exploited by personalized services to adapt the content to each user. For example, the collaborative filtering (CF) approach in recommender systems uses these preferences to suggest some items that comply with users' tastes and expectations.

Users' preferences can be expressed with text in natural language or with tags. They may also be expressed under the form of a *rating*-score on a multi-point scale. This modality has become one of the most popular ways of expressing one's preferences, specifically in e-services. The success of ratings is, in parts, due to the fact that they are automatically processable. They can be transformed into numerical values (if not yet numeric) and many operations can be conducted on them such as users' or items' average rating. Due to this facility, some algorithms have been designed to transform users' opinions expressed in natural language or with tags into a value point on the rating scale [14].

At the same time, ratings only require a small amount of time to express one's preferences. They are generally perceived as an easy task, but several works have

C. Huemer and T. Setzer (Eds.): EC-Web 2011, LNBIP 85, pp. 184–195, 2011.

highlighted important drawbacks, among which inconsistencies of ratings and limited precision [9,1]. We ask if we could not find another modality that would be as easy as ratings and that would not have ratings' drawbacks.

In this paper we propose a new modality based on the following acknowledgement: in everyday life, rating items is not such a natural mechanism. Indeed, we do not rate sweaters when we want to buy one. It is more likely that we will compare them two by two, and purchase the preferred one. Based on this observation, we propose to get users' overall preferences by asking them to *compare* pairs of items in place of asking them to rate them ("I prefer x to y").

In an exploratory work [3], we had shown that comparisons could be used as input preference data of a CF system and that, in some cases, the accuracy of the recommendations was comparable to that obtained with ratings. In this paper we focus on the relevance of this new modality, taking the user's point of view into account. We specifically concentrate on the way users perceive this modality, whether they express preferences similar to those revealed when rating items, and confront the quality of the recommendations deduced from each modality. After presenting the limitations of ratings, particularly in recommender systems, we focus on two user-studies we conducted to gather users' overall preferences on both expression modalities. We subsequently address three research questions: 1) Are users in favor of this new modality for expressing their preferences? 2) Is there a mapping between users' preferences expressed with both modalities? 3) Is the accuracy of the recommendations similar when they express their preferences by comparisons of items and by rating them? In this preliminary work, we discuss possible answers to these questions and show that asking users to compare items in place of rating them is a highly promising modality, especially for CF.

2 Related Work

2.1 Expressing Preferences with Ratings

Multi-point rating scales have become one of the most common feedback modalities. A great deal of research has studied the use of ratings as input preference data, but the fundamental issue of the optimal number of points in rating scales is still unresolved [12]. Several drawbacks have been identified: inconsistency, limited precision and influence of labels. We introduce these issues hereafter.

Inconsistency. Users' ratings face the well-known intra-user variability issue, also referred to as inconsistency or noise. This inconsistency may have several explanations. First, it is difficult to quantify one's preferences and to assign a rating point. Second, the mood and the context may influence the rating we assign to an item. Third, the granularity of scales may conduct to incoherences: if a scale is too large, users may have too many choices and assign different rating values to two equally liked items, or to one item at two different times [13]. When users are asked to rate items twice, their inconsistency has been evaluated at 40% [9]. A more recent work in recommenders showed that the noise in ratings may lead to a variation of the RMSE of more than 40% [1].

Limited precision. In many rating systems, the granularity of the scale is quite small, which may limit the discrimination power of ratings and thus their precision. A user might judge two items differently but give them the same score due to the limited scale. As a consequence, small scales may lead to imprecise ratings and possibly frustrated users [9]. In addition, although users' preferences are not linear (on a five-point rating scale there is a larger difference between a 2 and a 3 than between a 4 and a 5), the scales are processed as if they were, such as in CF; it may thus impact the quality of such systems [7].

Psychometric researches measured the reliability of different scales, with respect to granularity [12]. They showed that the less accurate scales (in terms of reliability, discriminating power and users' preferences) turn out to be those with the fewest number of rating points.

Maximal scale point. Because scales are bounded, once users have given the maximal rate to an item, they cannot express that any other item is better. This may have substantial consequences, as highly appreciated items are generally those that most reflect users' preferences. Recently, a first step towards the automatic customization of scales was achieved. Cena *et. al* showed that there is not always a bijection between two scales, confirming that their granularity influences the preferences expressed [7].

Influence of the meaning associated with the value points. It has been proven that, given a scale granularity, the values of the points and the descriptive labels associated with scale points have psychological influence on users, resulting in differences in the distribution of the ratings [2].

2.2 Expressing Preferences with Comparisons

The comparisons that we propose in this paper share some similarity with four feedback mechanisms, detailed in [11]. Whilst showing users an ideal item, they propose alternatives and use users' feedback to revise the current search query. *Value elicitation* and *tweaking* are two feature-level techniques, whereas *rating-based* and *preference-based* feedback methods operate at the case (or item) level. A popular version of the latter approach is *critiquing*, as proposed and studied by Pu and Chen [8]. A critique is for instance the feedback "I would like something cheaper than this item".

Despite these approaches relying on the act of comparing items, we are convinced that they are fundamentally different from our proposed *comparisons*, both in terms of goal and data representation. These feedback strategies are often directed at helping users to find an ideal product, and modelize the tradeoffs between compared items in terms of varying features (then used to update the query). The novelty of our paper resides in the fact that we aim to model users' overall preferences: preference-based feedback, not those corresponding to the current goal of the user, and above all that we record the preference relation between items, independently of items' attributes.

3 Motivation for Comparing Items

3.1 Advantages of Comparing

In Section 2 we showed that asking users to rate items in order to express their preferences has several drawbacks. However, few alternative preference expression modalities have become as popular as ratings. Reflecting on how we behave in real-life, where we often end-up comparing two items rather than rating them, we propose to use comparisons as a new modality for expressing preferences. Thus, rather than saying "I like this item and I give it a four out of five", a user will say "I prefer j to i" $(i < j)$, or "I prefer i to j"$(i > j)$, or "I appreciate them equally" $(i = j)$.

We believe that comparing items can be more appropriate than ratings for expressing preferences, for the following reasons:

- First, we are convinced that comparing items is easier than giving them a score. By asking users to compare items, the problem of quantifying a preference (Section 2) is avoided. In addition, [6] showed that making comparisons is faster than absolute judgement (ratings). We thus hope that using comparisons will lead to a higher users' participation rate.
- Second, we believe that comparing is less inconsistent than ratings as, contrary to rating, there is no need to remember previously compared pairs to compare a new one.
- Third, the problem of limited precision (Section 2) of ratings is avoided. When comparing items, users have a local point of view, focused on the two items to be compared. The resulting comparisons, represented as a preference relation [3], is made up of an un-predefined and adaptive number of levels.

One of the drawbacks of using comparisons is the increase in the number of comparisons needed to establish a ranking of items [5]. Another disadvantage is that no quantitative information is known about "how much" the user prefers an item to another. These issues are not the focus of this paper, and ways to alleviate them are discussed in [10].

Convinced that the advantages outweigh these drawbacks, we trust that comparing items can be more appropriate than ratings for expressing preferences.

3.2 Algorithmic Predisposition of Comparisons

In our recent preliminary work [3], we proposed a formalization of CF where input data is a set of comparisons. We showed that the classical memory-based approach can be used with such data. We also conducted experiments to evaluate the adequacy and the performance associated with the use of comparisons in CF. As we did not have any input data made up of "real" comparisons at our disposal, we simulated such a dataset. We used a corpus of ratings that we converted into comparisons. The resulting comparisons had thus the same drawbacks as ratings: inconsistency and limited precision. Furthermore, the quantitative dimension of ratings was lost during the transformation into comparisons. Even

so, the performance obtained was similar, and in some cases better to the one reached with ratings. We believe that these findings highlight the algorithmic predisposition of comparisons.

4 Experiments

4.1 Experiment Framework

We chose to work on the domain of motion pictures. We selected movies from the box-office, maximizing chances that users would be able to evaluate them. Our dataset was composed of 200 *films* and 150 *television series*. To run focused experiments, we built our own online website and relied on one evaluation page for each modality: *rating* or *comparing*. Both are shown in Figure 1 and displayed basic information (title of the movie, genre, year and poster). The rating page's feedback mechanism was a one to five star rating scale. The comparisons page displayed the same information but divided into two columns, A and B. Below both movies were three links that allowed users to express $a < b$, $a > b$ or $a = b$. For each movie, a large "I do not know this movie" button was available.

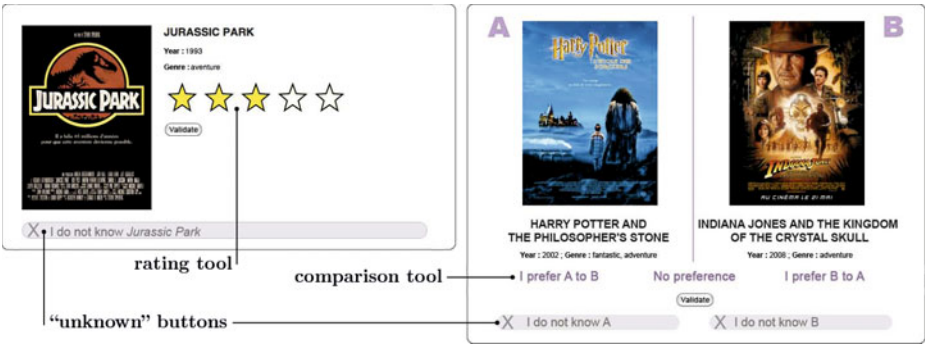

Fig. 1. The rating page (left) and comparison page (right) of the user-study

4.2 Evaluation Setup

We set up two user-studies. Experiment 1 sought to gather users' overall preferences between both expression modalities. We adopted a within-subject design: each user tried one modality (rating or comparing) on one dataset (films or tv series), before testing the opposite combination. Experiment 2 also relied on a within-subject design, but was more in-depth and aimed at understanding whether comparisons expressed the same preferences as ratings. For this reason, users first rated movies, and were then asked to compare pairs of the same movies the following day (the pairs of movies to be compared have been randomly selected within the set of rated movies). The one day gap was introduced to reduce the effects of learning and fatigue.

The general procedure in both studies was similar. Users received basic instructions, before starting a *three minute session* to either rate, or compare movies. Since a comparison concerns two items, rather than one for a rating, we decided to impose a fixed session duration. At the end of both sessions in Experiment 1, users were presented with three preference questions: `Q1` Which evaluation modality did you prefer? – `Q2` Which evaluation modality was the easiest to use? – `Q3` Which evaluation modality was the fastest to use?

As an incentive, EUR 10.- gift vouchers were proposed in a draw to users who had completed a study. Experiment 1 collected 100 users, with 52 males and 48 females. Users were mainly young (71% in the 18-24 age group), French (77%), familiar with Internet (98% use it daily) and watched films at least once a week (50%). We therefore expect them to be comfortable with the new proposed comparison modality. Experiment 2 being more detailed, only 25 users were recruited but their demographic distribution was similar.

5 Results

In this section we will first present the findings from Experiment 1, that focus on users' acceptance of comparisons. We then study the correspondence between preferences expressed through ratings and comparisons from Experiment 2. Last we focus on the quality of recommendations made when exploiting either users' ratings or comparisons.

5.1 Are Users in Favor of Comparisons?

In Experiment 1, after all participants had experienced both the rating and comparison mechanism once, we gave them a questionnaire asking each user to vote on which modality (rating, comparison or neither) they had liked most. Figure 2 shows the distribution of the 100 users' answers. With `Q1` we can observe that 53% of users preferred the comparisons, against 42% for the ratings. `Q2` indicates that 56% of users found comparisons to be easier than ratings. The amount of uncertain people is here higher, reaching 11%. Finally, for `Q3`, more participants found that it was faster to do comparisons than to rate, at respectively 54% against 42%. A Chi-square test of independence confirms that there was no ordering effect.

Overall, these results show that users are in favor of the comparison mechanism. Under all three tested dimensions, users found that the comparing modality was better than the traditional and wide spread rating mechanism. This is very encouraging: one must not forget that users have been confronted to rating systems for many years, not only online but also in real-life, especially on a topic such as movies. This was the first time they were confronted to a comparison mechanism.

5.2 Do Users Express the Same When Comparing and Rating?

In this section, we analyze the preferences expressed by the users in Experiment 2: when they compare items two by two versus when they rate items.

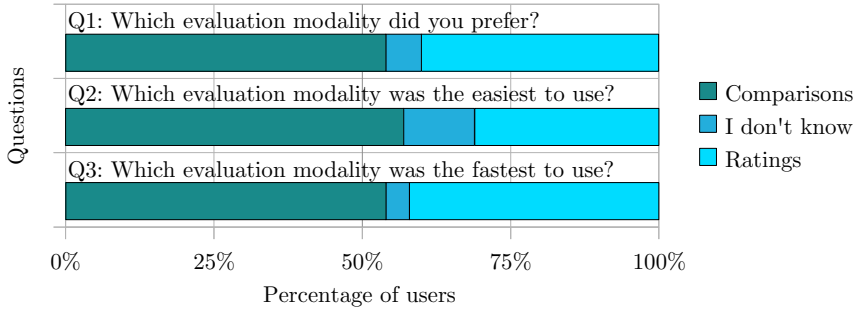

Fig. 2. Users' preferences between rating and comparison

The dataset containing the comparisons will be referred to as *CompDS* and the one containing ratings will be referred to as *RatDS*. Note that on average 33 ratings/comparisons have been collected within the three minute sessions.

A direct linking between *CompDS* and *RatDS* cannot be performed. Indeed, not only is the input data in *CompDS* made up of ordered pairs of items whereas it is single items in *RatDS*, but also the preference value is a comparison ($<$, $>$ or $=$) in *CompDS* and a rating score (from 1 to 5) in *RatDS*. Consequently we decided to transform one dataset into the format of the second. As comparisons contain no quantitative information, converting them into ratings is a challenging task. Oppositely, transforming ordered pairs of rated items into comparisons is straightforward: we chose to apply this conversion. For instance, if user u rated the item i_1 with a 5 and i_2 with a 4, this information will become the comparison $i_1 > i_2$. To allow a correspondence-computation between ratings and comparisons, not all pairs of items have been transformed into comparisons: we chose to transform only the pairs which had been compared by users in *CompDS*. The resulting corpus will be referred to as *RatCompDS*.

Table 1. Distribution of comparison values according to the preference modality

	Comparisons (*CompDS*)	Ratings (*RatCompDS*)
$i_1 < i_2$	42.4%	39.0%
$i_1 > i_2$	45.9%	38.5%
$i_1 = i_2$	11.7%	22.5%

Table 1 presents the proportion of each comparison value ($<$, $>$ or $=$), for both modalities. First, we can see that the distribution of $<$ and $>$ is homogeneous in both modalities. Second, users assign identical ratings to pairs of items in 22.5% of the cases. However this is around twice more than the percentage of cases where they consider items as equivalent (11.7%) when they compare them.

Table 2 details the correspondence between preferences expressed in *CompDS* and those in *RatCompDS*. Each line of the table represents one comparison

Table 2. Correspondence of rating preferences and comparison preferences

		Ratings (*RatCompDS*)		
		$r(i_1) < r(i_2)$	$r(i_1) > r(i_2)$	$r(i_1) = r(i_2)$
Comparisons (*CompDS*)	$i_1 < i_2$	**74.1**	6.1	*19.8*
	$i_1 > i_2$	8.9	**71.3**	*19.8*
	$i_1 = i_2$	30.1	27.2	42.7

value in *CompDS*. They show the distribution of the ratings on the corresponding pairs of items in *RatCompDS*, and sum up to 100%.

When users compare two items and judge them as *different*: $i_1 < i_2$ or $i_1 > i_2$, the corresponding ratings have the same trend in respectively 74.1% and 71.3% (on average 72.7%) of the cases. In the remaining 27.3%, 19.8% correspond to equal ratings. This means that although users judge two items as being different through a comparison, they assign them both the same rating. This can be explained by the limited precision of ratings highlighted in Section 2. For this reason, we believe that it is reasonable to consider these 19.8% as non contradictory preferences. Consequently, we can say that when users judge two items as different through comparisons, in 92.5% of the cases they assiged coherent ratings: equal ratings or ratings with the same trend.

When users compare i_1 and i_2 and judge them as *equivalent*, they give them the same ratings in only 42.7% of the cases. When focusing on the 57.3% of remaining cases, 42% correspond to pairs of adjacent ratings (that differ by only 1 point). This high value may be explained by the inconsistencies of users' ratings presented in Section 2, and by the fact that no precise meaning had been associated to each rating value in the experiments. Thus, these 42% should be considered as coherent with the comparisons. Consequently, we believe that when users compare two items as being equivalent, the ratings are coherent with these comparisons in 84.7% of the cases: they assign them similar or adjacent ratings.

As a conclusion, we feel that it is reasonable to say that, although there is no direct mapping between ratings and comparisons, they are mainly coherent.

We conducted an additional evaluation, with the aim of studying the pertinence of exploiting comparisons in the frame of CF. We raise the following question: are the respective top-n (preferred) items the same in *RatDS* and *CompDS*? For each user u, we build the preference relation that corresponds to his/her comparisons (as done in [3]). The number of ranks of these preference relations varies according to the users, from 3 to 9 levels. We then ask if the items on the top ranks in the preference relations are the preferred items in terms of ratings? Figure 3 presents the distribution of the ratings according to the first three top-rank values in the preference relation. We can see that the items on the top of the preference relations (rank 1) are mainly items with rating values of 5 and 4 (average rating: 3.99). When the rank of the items increases, the average rating value decreases. The average rating of items in rank 2 is 3.07 and the one in rank 3 is 2.60. The graph supports that items highly ranked in the preference relation extracted from comparisons, tend to be those that have been preferred by users in the sense of ratings.

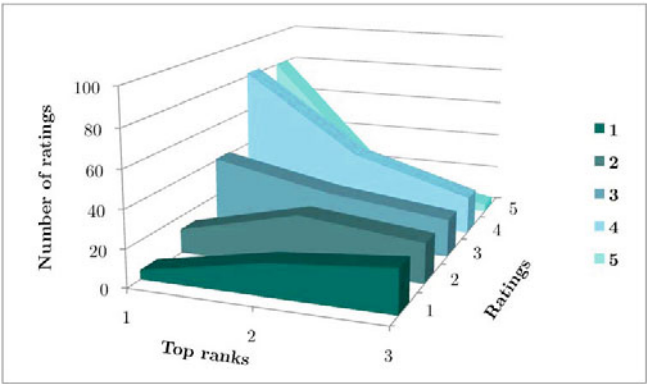

Fig. 3. Repartition of ratings in function of top ranks in preference relations

5.3 How Accurate Are the Recommendations from Both Modalities?

To obtain an initial impression of the potential of comparisons, we conducted a small-scale experiment in the frame of CF. We asked users to evaluate the quality of the recommendations generated with either of the two preference expression modalities. To build a recommendation list from ratings, we used a classical memory-based CF approach with the cosine measure as similarity between users [4], computed on users' preferences acquired from previous experiments. To build a recommendation list from comparisons, we used the above memory-based collaborative filtering, adapted to comparisons, as was already done in [3]. As we used the same recommendation algorithm in both cases, the quality of the recommendation lists are directly comparable. We built a recommendation list for only the 25 users from the second study; all the users from the first study were used to compute users' similarities.

First, we asked each user to rate the top 10 items from the recommendation list, they could rate (whether they had seen them or not). To ensure that each user can rate 10 items (and that the resulting rating lists are comparable) we presented recommendation lists of 30 ordered movies (starting from the best). Figure 4 presents, for each rating value, the average number of items that have been rated with this value, in each rated list. We can see that the average number of items which received a top rating value (5 and 4), is larger in the comparisons' lists. The ratings' lists contain more low rating values. The distribution of comparisons is centered around higher grades than for ratings.

Second, we collected users' global opinion on the recommendation lists by asking them which one they preferred. 16 users preferred the recommendations from comparisons against 9 for ratings. Without trying to read too far into these results[1], we can confidently say that our proof-of-concept worked: the

[1] Due to the small number of users, statistical tests could not be computed.

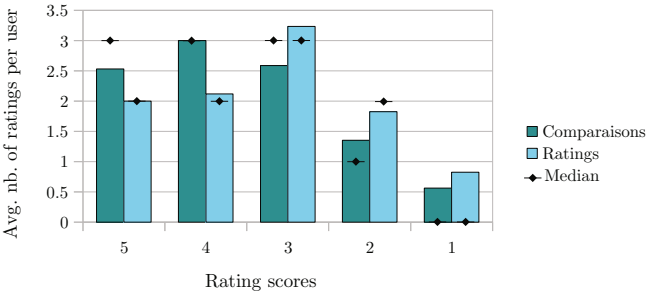

Fig. 4. Distribution of rating scores across the top-10 recommendations

comparisons appear to have generated recommendations at least as valuable to users, as a rating based approach. This finding confirms our exploratory work [3].

6 Discussion

Our results show that when focusing on pairs of items, comparisons are mainly coherent with ratings (in more than 85% of the comparisons). From a global point of view, the analysis of the preference relation versus the set of ratings also shows similar tendencies. Nevertheless, the two modalities are not equivalent and would deserve a more refined analysis. When we constructed the preference relations, we only made sure that all items were compared and connected in one single graph. Unfortunately, some relations (comparisons) are of high importance to build an accurate preference relation, whereas others can be useless. For instance, supposing we know that $i > j$, finding out that $k > j$ says little about the relation between i and k, whereas knowing that $k > i$ allows to propose that $k > i > j$ (in case we assume transitivity). Even though this issue has no consequence on the analysis of pairs of items, it influences the global perspective. Thus we believe that the comparative examination of both modalities could be refined by controlling which comparisons are presented to users [10].

The findings also reveal that the comparison modality solves the problem for choosing the optimal rating scale. Indeed, when asking users to compare items, they unconsciously build their own scale, with the granularity that fits their preferences. We observed that for some users, three levels are enough, whereas others need up to nine levels of ranking (within the three-minute timeframe). We are therefore confident that comparisons can be an excellent answer to the problem of customizing rating scales, raised in [7].

In the case of inconsistencies in comparisons, the task of de-noising preferences is facilitated. Indeed, the relation between two items can be known or deduced from several relations in the preference relation. Thus, in the case of inconsistencies in preferences, the choice of the edges to be kept is facilitated (for example by using a majority vote).

Our results showed that, although quite similar to ratings, comparisons seem to allow users to express finer preferences, especially when users' ratings were

equal. However, reflecting on long-term perspectives, we do not yet envisage to solely exploit preferences acquired through comparisons. Because of the qualitative nature of comparisons, it is possible to have a preference relation, made up of several levels, where the top item may still not be liked by the user. When exploiting comparisons in CF, the knowledge of items that have actually been liked is crucial so as to not recommend items that users would not like. At the same time, as the number of levels in the preference relation grows, this quantitative problem disappears. Consequently, some absolute preferences (such as ratings), might be useful to ensure the accuracy of recommendations at first, and we envisage to hybridize both modalities in our future work.

We believe that we could exploit ratings to establish a first classification, before refining the highly rated items by using comparisons. However, we cannot envisage to ask users to express their preferences with both modalities. To solve this problem, we could collect users' implicit feedback from which we could deduce ratings, viewed as an additional information to comparisons. We could also use this deduced information to identify appreciated items and refine by asking users to compare them.

7 Conclusion

The most popular modality for expressing one's preferences is rating: on a predefined multi-point scale, we choose the point that reflects best our preference. However, although several studies have put forward drawbacks of ratings (inconsistency, limited precision, etc.), no other modality has yet supplanted ratings. We have proposed an alternative: comparisons, that asks users to compare pairs of items. To assess the pertinence of this modality, we performed two user-studies. We show that users are in favor of comparisons as they find them easier, faster and on the whole prefer them. Our results also reveal that comparisons express preferences similar to those of ratings, as ranks in preference relations seem to be coherent with ratings. To finish this inital work, we generate recommendations based on either ratings or comparisons, and show that comparisons give very promising results. Consequently, we are convinced that comparisons are a highly promising new modality for preference expression, that could possibly improve the user experience, especially in the frame of collaborative filtering.

These initial findings encourage us to explore comparisons in depth. We are studying the stability of comparisons through time *vs.* that of ratings. To cope with the problem of the large number of comparisons required, we focus on a strategy about the sequences of comparisons to be asked to users, in order to build a precise preference relation while minimizing the number of comparisons asked to users.

References

1. Amatriain, X., Pujol, J.M., Oliver, N.: I like it.. I like it not: Evaluating User Ratings Noise in Recommender Systems. In: Proc. of UMAP Conf. (2009)
2. Amoo, T., Friedman, H.: Do numeric values influence subjects' responses to rating scales? J. 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 at ECML-PKDD (2010)
4. Candillier, L., Meyer, F., Boullé, M.: Comparing state-of-the-art collaborative filtering systems. In: Proc. of 5th International Conference on Machine Learning and Data Mining in Pattern Recognition (MLMD), pp. 548–562 (2007)
5. Carterette, B., Bennett, P.: Evaluation measures for preference judgments. In: Proc. of the Annual ACM SIGIR Conference, pp. 685–686 (2008)
6. Carterette, B., Bennett, P., Chickering, D., Dumais, S.: Here or there; preference judgments for relevance. In: Proc. of ECIR, pp. 16–27 (2008)
7. Cena, F., Vernero, F., Gena, C.: Towards a customization of rating scales in adaptive systems. In: Proc. of the User Modeling, Adaptation, and Personalization (UMAP), pp. 369–374 (2010)
8. Chen, L., Pu, P.: Experiments on the preference-based organization interface in recommender systems. ACM Trans. Comput.-Hum. Interact 17, 5:1–5:33 (2010)
9. Cosley, D., Lam, S., Albert, I., Konstan, J., Riedl, J.: Is seeing believing?: how recommender system interfaces affect users' opinions. In: Proc. of the Intelligent User Interfaces (IUI) (2003)
10. Jones, N., Brun, A., Boyer, A.: Initial Perspectives From Preferences Expressed Through Comparisons. In: Int. Conf. on Human-Computer Interaction (July 2011)
11. McGinty, L., Smyth, B.: Comparison-based recommendation. In: Craw, S., Preece, A.D. (eds.) ECCBR 2002. LNCS (LNAI), vol. 2416, pp. 575–589. Springer, Heidelberg (2002)
12. 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)
13. Schafer, B., Frankowski, D., Herlocker, J., Sen, S.: Collaborative Filtering Recommender Systems. In: The Adaptive Web. Methods and Strategies of Web Personalization, pp. 291–324. Springer, Heidelberg (2007)
14. Turney, P.: Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. In: Proc. of the Association for Computational Linguistics (ACL), pp. 417–424 (2002)