User-controllable Personalization: A Case Study with SetFusion

Denis Parra
Computer Science Department
Pontificia Universidad Catolica de Chile
Avenida Vicuña Mackenna 4860
Santiago, Chile
dparra@ing.puc.cl

Peter Brusilovsky
School of Information Sciences
University of Pittsburgh
135 North Bellefield Ave., Pittsburgh, PA 15260, USA
peterb@pitt.edu

Corresponding Author:
Denis Parra
Computer Science Department
Pontificia Universidad Católica de Chile
Avenida Vicuña Mackenna 4860
Santiago, Chile
dparra@ing.puc.cl
+56 (2) 2354-4442

Abstract.

In this research we investigated the role of user controllability on personalized systems by implementing and studying a novel interactive recommender interface, SetFusion. We examined whether allowing the user to control the process of fusing or integrating different algorithms (i.e., different sources of relevance) resulted in increased engagement and a better user experience. The essential contribution of this research stems from the results of a user study (N=40) of controllability in a scenario where users could fuse different recommendation approaches, with the possibility of inspecting and filtering the items recommended. First, we introduce an interactive Venn diagram visualization, which combined with sliders, can provide an efficient visual paradigm for information filtering. Second, we provide a three-fold evaluation of the user experience: objective metrics, subjective user perception, and behavioral measures. Through the analysis of these metrics, we confirmed results from recent studies, such as the effect of trusting propensity on accepting the recommendations and also unveiled the importance of features such as being a native speaker. Our results present several implications for the design and implementation of user-controllable personalized systems.

Keywords Recommender Systems, User Studies, Interactive User Interfaces, Hybrid Recommender System, User-centric evaluation.

1. Introduction

The purpose of recommender systems is helping a user or a group of users to choose items from a large item or information space (McNee, Riedl, & Konstan, 2006a) by proactively suggesting personalized relevant items. Recommender systems were introduced in the early 90s with systems like Tapestry for filtering e-mails (Goldberg, Nichols, Oki, & Terry, 1992), GroupLens for netnews recommendations (Resnick, Iacovou, Suchak, Bergstrom, & Riedl, 1994), or Ringo for music recommendation (Shardanand & Maes, 1995), and several factors have helped to increase their popularity over time. For one thing, the exponential growth of the Internet makes it an ideal "large information space" to create recommendations for several applications and domains, such as the product recommendation of e-commerce websites like Amazon.com, the movie recommendations of Netflix, or the video recommendations of the web portal YouTube. Another factor that has popularized recommenders in areas beyond their original niches has been online open competitions such as the "Netflix Prize" (Bennett, Lanning, & Netflix, 2007) –a movie recommendation challenge that awarded one million dollars to the most accurate recommendation approach. Despite their success, recommender systems also face several challenges. One such challenge is incorporating Human Factors in order to increase user acceptance of the systems and the items recommended. Historically, the focus on recommender systems' research has been on improving the algorithms' predictive accuracy (Parra & Sahebi, 2013), but as McNee et al. (2006b) highlighted in the paper "*Being accurate is not enough: how accuracy metrics have hurt recommender systems,*" accuracy does not always correlate with a good user experience, making the study of recommender interfaces one of the areas in need for improvement.

The work on increasing user acceptance of recommender systems through better recommendation interfaces started with the exploration of visually-rich recommendation interfaces that go beyond the paradigm of static ranked lists. PeerChooser (O'Donovan, Smyth, Gretarsson, Bostandjiev, & Höllerer, 2008), and SmallWorlds (Gretarsson, O'Donovan, Bostandjiev, Hall, & Höllerer, 2010) are examples of interactive visual interfaces that represent a collaborative filtering paradigm, where users increased their satisfaction under the visual interactive interface compared to a more static condition. More recent work has focused on providing users control over the recommendation interface by allowing users to sort the recommendation list based on different item features in an energy-saving application (Knijnenburg, Reijmer, & Willemsen, 2011), by letting users indicate their preferences at different levels of granularity in a music recommender (Hijikata, Kai, & Nishida, 2012), or by permitting them to combine several recommendation sources using sliders in a music and a job recommender (Bostandjiev, O'Donovan, & Höllerer, 2012, 2013). These approaches have shown in particular domains how user controllability and user characteristics affect the user acceptance of recommendations.

Our work extends past research on both visual recommender interfaces and user controllability in two important directions. First, it suggests a new approach to user-controllable hybrid recommendation that combines more traditional *sliders* with a new way to inspect and control a fusion of recommendations through a *Venn diagram visualization*, inspired by our recent results in (Verbert, Parra, Brusilovsky, & Duval, 2013). Second, it examines the effect of controllability and user characteristics on the user experience by using objective, subjective and behavioral measures. This second contribution helps to bridge the gap of previous studies that consider only objective, only subjective, or at most both types of metrics, but that do not explain the user experience by describing how users interact with available widgets.

In order to address these challenges, we have built a novel user-controllable article recommendation interface for the existent system Conference Navigator, an online web platform that supports attendees and organizers of academic conferences. Using this system, we have conducted a controlled user study to investigate the effect of user controllability on the user experience of a personalized recommender system. The study compares two interfaces, a traditional static list of recommendations (baseline) against a visual interface with controllable features, and also investigates the effect of users' characteristics on the acceptance of recommendations.

The rest of the paper is structured as follows: in section two we survey previous work that motivates and influences our research; then in section three we present in detail the innovations of our work with a focus on our interactive interface. The design of our user study is described in section four, including research questions and related metrics, recommendation approaches and the study procedure. The results of our study are split into three sections. Section five presents how people used particular features of the interface (sliders, filters through Venn diagram) and how effectively those features improved user engagement, then section six aggregates the results of three sets of regression analyses in order to understand the influence of the interface and user characteristics on objective, subjective and behavioral metrics. Following, section seven shows the results of the post-study survey with a qualitative analysis of user comments. Finally, section eight summarizes the main lessons and conclusions of our research and it describes our future work.

2. Related Work

In this section we present the previous work that motivates our research. Since the areas described might have been explored in different research fields, we focus mainly on summarizing the work directly related to personalization and recommender systems. Hence, we classify the most relevant related work in three areas: a) Controllability, inspectability and user intervention, c) Transparency and explainability, d) User-centric evaluation of recommender systems.

2.1. Control, Inspectability and User Intervention

Though there are existent works on the effect of increased user control in online systems (Ariely, 2000), (Sherman & Shortliffe, 1993), only recently has user-control been methodically investigated in the context of recommender and personalized systems. Jameson et al. (2006) studied how much control users prefer on updating a list of recommendations in the context of a conference (UM 2001), but they did not find conclusive preference for one condition over another, but instead discovered several situational and individual factors that might affect the user's preference. Knijnenburg et al. (2011) studied the effect of different interaction mechanisms on an energy-saving recommender system. They concluded that the best interaction mechanisms depend on user characteristics; for instance, expert users (with more domain knowledge) reported higher user satisfaction with interfaces that provided more control compared to novice users, who were more satisfied with an interface that provided the recommendation without many controllable variables. Bostandjiev et al. (2012) introduced a visual hybrid interactive music recommender called TasteWeights and they performed a study to see whether the additional interaction results in a better user experience. Recommendation accuracy, measured as the utility of the recommended list of items after users have "tuned" the importance of different data sources and neighbors using the visualization, was better with bigger interaction and explainability (the full interface), as was the general user experience. Using the same TasteWeights framework, but only considering social recommendation (Facebook contacts) of music, Knijnenburg et al. (2012) performed

a user study on the influence of control and inspectability on the user experience. Letting users inspect the full recommendation graph (items, friends, and connections), produced an overall better user experience. In terms of type of control, they conclude that controlling weights and controlling the weight of items are additive, so providing both in a real setting is recommended. Hijikata et al. (2012) also explored control in the context of music recommendation. They explored four different ways to let users intervene the recommendation process: ratings, context, content attribute and user profile edition. Through a user study, they showed that user intervention is correlated to rating prediction and user satisfaction, but user control doesn't always lead to better prediction and satisfaction. In addition, they found preliminary evidence that only people with high interest in the domain consistently experience better user satisfaction with more control, even when recommendations are less accurate. Another approach to user controllability was studied in (Verbert et al., 2013), where an interactive interface was embedded into a conference support system, a visual interactive tool called Aduna[1]. This tool was adapted to aid users in exploring talks in a conference from multiple perspectives of relevance-talks bookmarked by users, suggestions of recommender agents and talks marked with specific tags.

2.2. Transparency and Explainability in Recommender Systems

(Herlocker, Konstan, & Riedl, 2000) introduced the idea of explaining recommendations as a mean to make the system more transparent to users' decisions and to improve users' acceptance of recommender systems. Based on successful previous results from expert systems, they expected that interfaces of collaborative filtering recommenders would benefit from explanations as well. They studied different ways to explain recommendations and rated histograms "the most compelling way to explain the data behind the prediction." A study with 210 users of MovieLens, a well-known movie recommender system, showed that users value explanations and would like to add them to the recommender interface (86% of the respondents of a survey). The authors also think that explanation facilities can increase the filtering performance of recommender systems, though they couldn't find explicit evidence to support it and called for further well-controlled studies in this area. Furthermore, (Tintarev & Masthoff, 2007) notice that explanations might have different objectives, and identify seven different aims for explanations: transparency, scrutability, trustworthiness, effectiveness, persuasiveness, efficiency and satisfaction. More recently, in the handbook of recommender systems there is a whole chapter that addresses the design and evaluation of explanations in recommender systems (Tintarev & Masthoff, 2011).

2.3. User-centric Evaluation of Recommender Systems

Traditionally, evaluation of recommender systems has relied mainly on prediction accuracy, but over the years researchers and professionals implementing recommender systems have reached consensus that this evaluation must consider additional measures such as diversity, novelty, and coverage, among others. Beyond these metrics, recent research has increasingly considered user-centric evaluation measures such as perceived diversity, controllability and explainability. For instance, Ziegler et al. (2005) studied the effect of diversification in lists of recommended items, Tintarev and Masthoff (2007) investigated on recommender systems' transparency, Cramer et al. (2008) studied explainability in

[1] http://www.aduna-software.com/technology/clustermap

recommender systems, and Knijnenburg et al. (2012) tried to explain the effects of user-controllability on the user experience in a recommender system.

Nevertheless, as a result of a lack of a unified framework, comparing the results of different studies or replicating them is not a simple task. Two recent user-centric evaluation frameworks addressed this issue. On one side, Pu et al. (2011) proposed ResQue, identifying four main dimensions (perceived quality, user beliefs, user attitudes and behavioral intentions) and a set of constructs to evaluate each one. On the other side, Knijnenburg et al. (2012) defined dimensions and relations between them (objective systems aspects, subjective system aspects, experience, interaction, situational characteristics and personal characteristics), but encouraged the users of this framework to choose their own constructs based on some specified guidelines.

2.4. SetFusion in the Context of Related Work

Compared to the aforementioned studies, our research contributes to user-controlled personalization by implementing a new way to visualize and filter a group of recommendations through an interactive Venn diagram and also by studying behavioral patterns that can indicate why an interactive controllable interface might increase or decrease the user's acceptance of the recommendations. Considering in our assessment the system's performance in conjunction with user-centric aspects, we conducted a three-fold approach: objective, subjective and behavioral measures. In this ongoing effort, we have already presented two short studies: one with a preliminary version of the user-controllable interface in the context of the CSCW 2013 conference (Parra & Brusilovsky, 2013), and an updated version of the system in a small-scaled field study in the context of the UMAP 2013 conference (Parra, Brusilovsky, & Trattner, 2014). These studies allowed us to obtain hints about the potential of our proposed interface and alternatives to improve it, but the small number of subjects prevented us from testing specific hypotheses and the study setting did not allow us to control for the effect of user characteristics, topics that are addressed in this paper.

3. SetFusion Visual Recommender Interface

The main innovation implemented in our research is a user-controllable transparent hybrid recommendation interface that we call SetFusion. In order to evaluate our ideas of user-controllable recommendation, we integrated SetFusion into a conference support system, Conference Navigator 3 (CN3). CN3 (Parra, Jeng, Brusilovsky, López, & Sahebi, 2012) supports conference attendees by providing traditional conference information (i.e., conference program, proceedings, list of participants) enhanced with paper recommendations and social navigation features like a list of popular bookmarked talks. By the time of writing this article, CN3 has been used to support over 24 conferences. The decision to use Conference Navigator to conduct the user study was based on two main factors: the amount of conference data stored by the system to produce recommendations (detailed conference proceedings and users' bookmarked talks), and the user-tracking functionality available through CN3 API, which is critical to analyze users' behavior. Since CN3 uses different sources of information to generate paper recommendations (papers' popularity by user bookmarking, content-similarity, author reputation, etc.) a hybrid recommender system is ideal since it can integrate several recommendation strategies (Burke, 2002). In contrast to traditional hybrid recommendation approaches where the fusion of sources is done behind the stage and users see the traditional ranked list, SetFusion implements a user-controllable and transparent visual hybrid recommendation interface. The following explains in detail the design of SetFusion and its components.

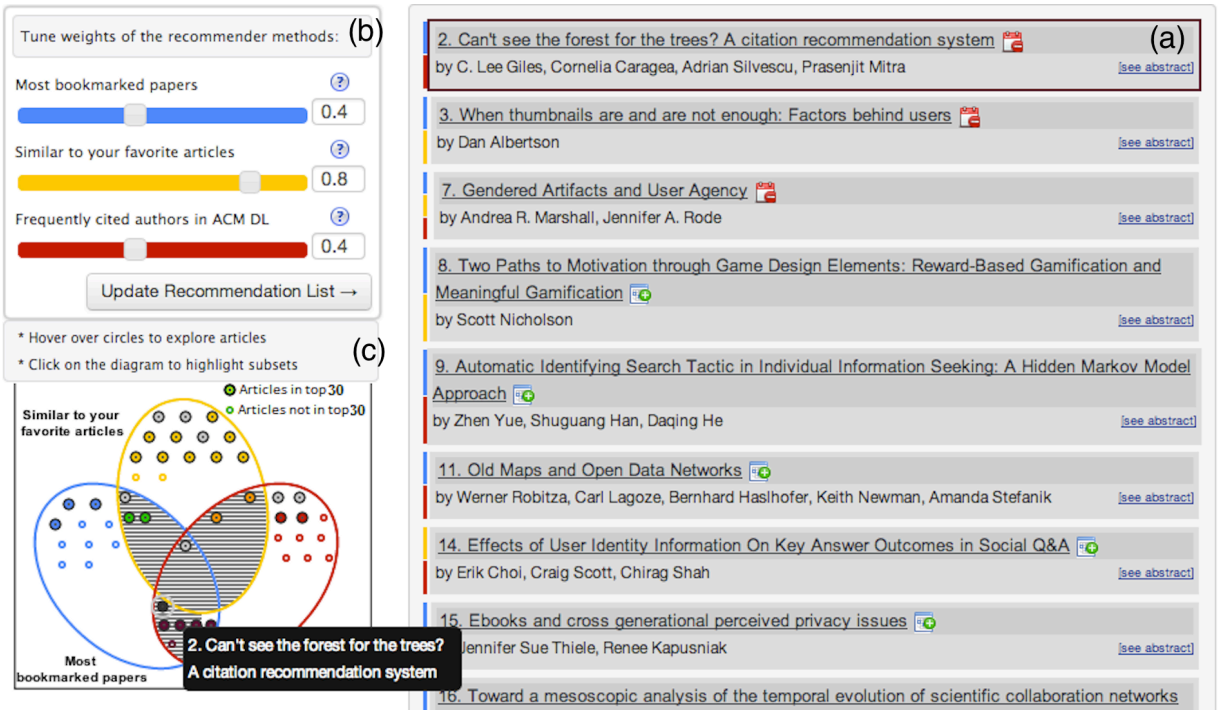


Figure 1. Screenshot of SetFusion interface. The three components indicated are: (a) list of recommendations, (b) sliders, and (c) Venn diagram.

The SetFusion interface (Figure 1) consists of three main components: (a) list of recommendations, (b) sliders to adjust the importance of each recommendation method, and (c) interactive Venn diagram.

The list of recommended items (Figure 2) presents recommended papers ranked by relevance, with the most relevant at the top. A color bar on the left side of each paper indicates the method(s) that recommended the paper. The list supports four tracked actions:

a) Open and Close Abstract: by clicking on the link provided next to each paper title, the users could see the abstract of the paper.
b) Hover over color bar: hovering over the color bar brings up an explanation of the method used to recommend the paper.
c) Bookmark a paper: at the very end of each paper's title, a red/green icon indicates if the paper is bookmarked or not, allowing also to add or remove the paper to/from the list of user bookmarks.
d) See 10 more: By default, the system shows the top 30 recommended items. If the user wants to see more items below that point, she can click on the button "See 10 more." This button allows registering cases where the top 30 items were not sufficient to find interesting talks.

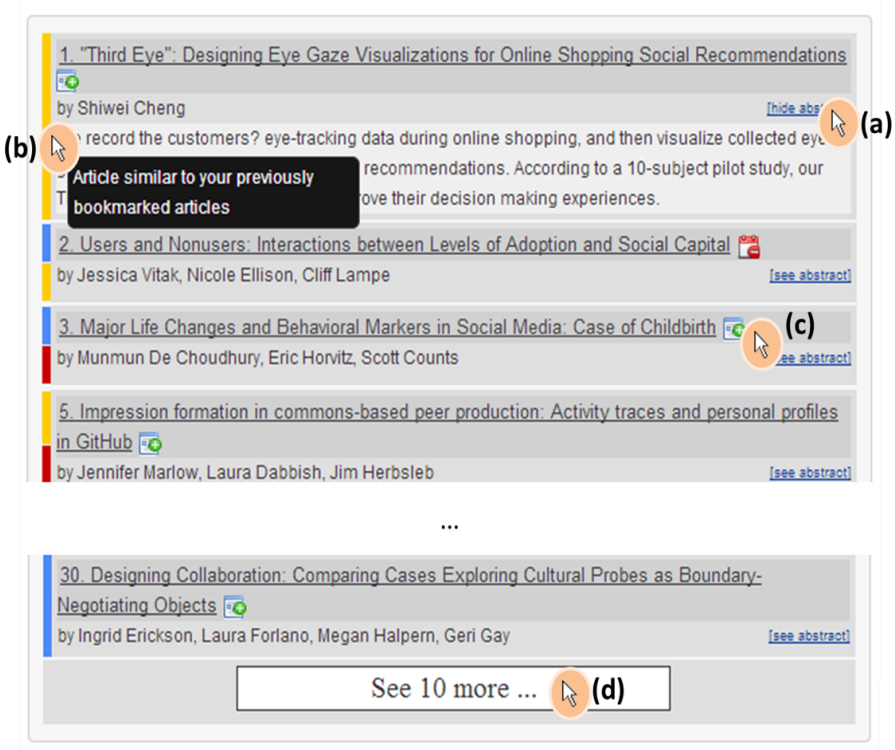


Figure 2. Screenshot of the recommended items list. The arrows highlight the actions that the user can perform in the recommender interface.

The Sliders (Figure 3) are the key to the controllable fusion of recommendation. Each slider corresponds to a specific color-coded recommendation method. The slider position represents the importance currently assigned to method. Working with the sliders, users can:

a) Hover over the help icon to obtain a more detailed explanation of the method.
b) Move sliders to change the relative importance of each method used to generate the list.
c) Re-generate the fused recommendation list according to the current position of sliders by clicking on the button "Update Recommendation List".

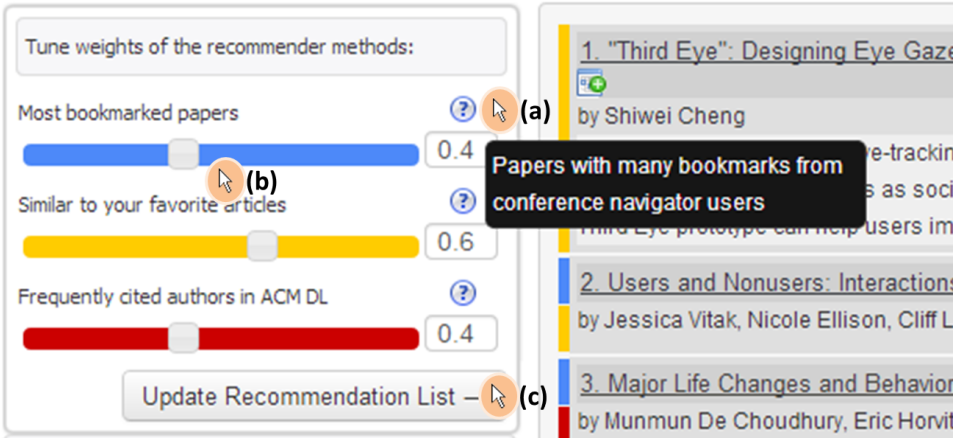


Figure 3. Screenshot of the sliders widget. The arrows highlight the actions that the user could perform in the controllable interface.

The Venn diagram (Figure 4) provides a set-based representation of the items recommended by each method. Here each color-coded ellipse represents a recommendation method and each small circle represents one of the recommended talks (also shown in the ranked list). The position of each circle indicates which method(s) recommended the corresponding talk. Talks located on the "intersections" are recommended by more than one method. The actions available on this widget are:

a) Hover over the talk circle to open a floating dialog with the title of the talk.
b) Click on the talk circle to *scroll* the ranked list on the right panel to focus on the talk.
c) Click on a Venn diagram ellipse or intersection area to *filter* the list on the right panel (c-2) so that it only shows the articles located in on the clicked area, i.e., recommended by the method(s) represented by the ellipse or intersection of ellipses.

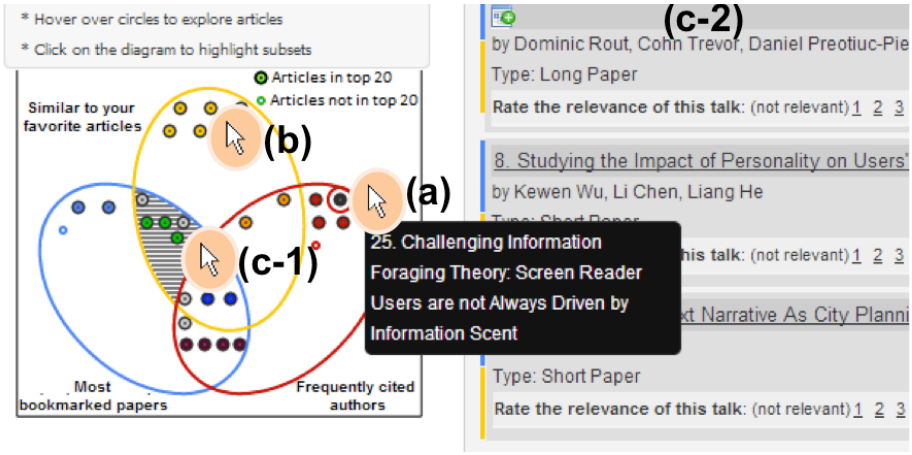


Figure 4. Screenshot that highlights the actions available on the Venn diagram widget.

This interface implements several controllability and transparency aspects: (a) the user can control the fusion coefficients and corresponded ranking using sliders; (b) the user can control filtering using the Venn diagram; (c) the current position of the sliders and highlighted area on the Venn diagram clearly show how the current ranked and filtered list was generated; (d) color bar on the left of each talk and the position of the talk circle on the Venn Diagram explains why this talk was recommended; (e) several kinds of inspectability are supported through hovering on the Venn diagram and other interface components.

4. Study Design

In order to investigate the effects of our approach to user controllability, we designed a user study that tested the user experience with SetFusion compared to a baseline interface. The task consisted of asking people to use Conference Navigator to choose interesting papers simulating a conference preparation scenario some days before a conference takes place. The study had a within-subjects design and its workflow is presented in Figure 5. A subject started by completing a pre-survey and then was assigned to one of two possible sequences of interfaces to perform the *Bookmarking* task. If the subject was assigned to the sequence at the top, we first captured her preferences (*Pref* step in Figure 5), then she proceeded to use the baseline non-controllable (*No-C*) interface for recommended papers, and then continued with the controllable (*C*) interface.

The *Pref* step consists of asking the subject to examine all papers from the proceedings of an *iConference* series conference (in the order of 50 papers showing title, author(s) and abstract) and bookmark the ones that she finds relevant, without a limit in the number of papers to be bookmarked.

Then she proceeds with the two rounds of bookmarking (first with *C* and second with *No-C* interface). *Each* round simulates a realistic scenario of preparing for a conference by finding and bookmarking interesting talks to attend. In each round, we use full data from the *iConference* series. Note that different conferences are used in *No-C* and *C* rounds of bookmarking and the *Pref* step. In total, three conferences from this series (*iConference 2011*, *iConference 2012*, and *iConference 2013*) have been used in the study. After working with each version of the interface, the subject is requested to complete a post-session survey that captures the perception of the interface just used. Finally, after bookmarking papers using both interfaces, the subject is asked to complete the *Rating* task, where she rates (on a scale from 1 to 5) the relevance of *all* papers she saw in the *Bookmarking* task (i.e., from all three conferences). This exhaustive rating task is placed at the end to avoid disrupting natural bookmarking behavior in the *Bookmarking* task. At the very end, the user completed a post-study survey with the purpose of comparing her impression about both recommender interfaces and obtaining additional user feedback.

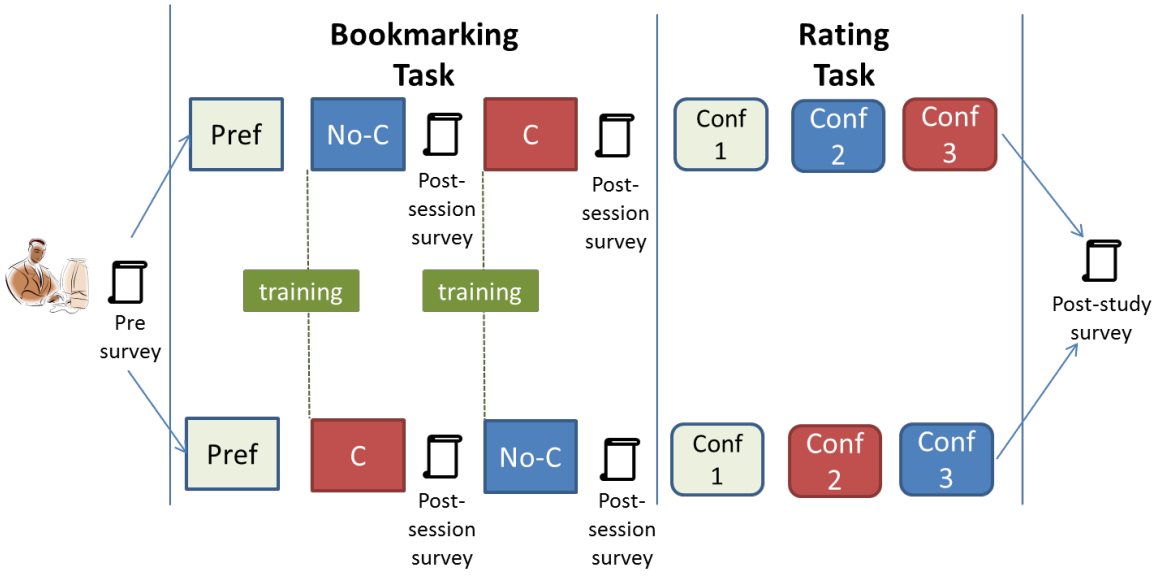


Figure 5. Workflow of user study. After answering the entry questionnaire (pre-survey), the subject was assigned to one of 2 possible sequences of interfaces: Non-controllable (No-C) and then controllable (C) interface, or vice versa.

4.1. Baseline Interface

In order to test the hypotheses about the SetFusion interface, we compared it to a baseline which should resemble traditional recommender interfaces. In terms of research design, we considered SetFusion our treatment condition, and the baseline interface the control condition. Figure 6 presents the baseline interface that we designed for the study. It resembles the SetFusion interface, but the sliders and Venn diagram widgets have been removed. We also removed the color bars at the left side of each talk, used in SetFusion to indicate a method or a combination of methods that recommend the talk. In this baseline interface, we are tracking the following actions: (a) Bookmark: when the user finds an item considered relevant, she clicks in the icon "bookmark this item;" (b) See abstract: by default, papers are presented only by their title and authors, but they can click a button to see the article's abstract; and (c) See 10 more: by default we display the 30 top ranked recommended items, but users can scroll down the list, as in Figure 7, if they still have not found enough relevant papers.

Figure 6. Screenshot of the baseline interface. The interface presents a static list of conference talks, which was built using a hybrid recommendation method.

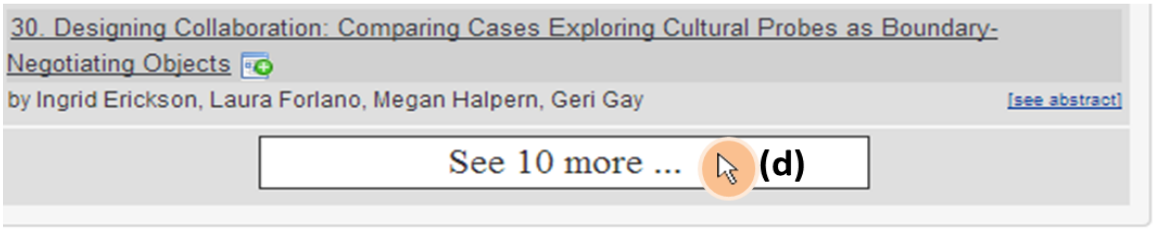


Figure 7. By default, the list of recommendations shows 30 articles. The button "See 10 more" allows users to expand the list of recommended items, as many times as items recommended are available.

4.2. Research Questions

Our general expectation was that SetFusion, which provides the user control and inspectability, would induce an increased engagement and a better user experience. However, due to the results of previous related work, we also expected that some user characteristics would affect how users accept recommendations being provided with more or less controllability. We formalized our expectations and assumptions in the following research questions, with details on our motivation to address them:

RQ1. How does controllability affect user engagement with a recommender system?

This question was motivated by an existent gap between the research on user engagement in software applications that highly recommends the use of both subjective and objective metrics (O'Brien & Toms, 2010) and the results of previous research (Hijikata et al., 2012; Knijnenburg, Bostandjiev, et al., 2012), that support the effect of controllability on user engagement, mainly by considering subjective measures (surveys) and one or very few objective metrics (average rating). We attempted to bridge this gap by studying controllability and evaluating user engagement more comprehensively than previous research.

RQ2. How does controllability affect the user experience in a recommender system?

The motivation of this question is similar to the previous one, but with a more holistic perspective. The user experience involves user engagement, but also many other dimensions such as perceived system quality, user beliefs, user attitudes and behavioral intentions, that have been recently formalized in evaluation frameworks proposed by Pu et al. (2011) and Knijnenburg et al. (2012). Though we do not present an evaluation using strictly the aforementioned frameworks, we consider and adapt all their dimensions to assess the user experience with subjective metrics. Moreover, we also consider objective metrics and behavioral measures to complement the evaluation, thus implementing a comprehensive evaluation.

RQ3. Do user characteristics affect the influence of controllability on the user's engagement with a recommender system?

Previous research has shown that user characteristics affect engagement with online systems. For instance, O'Brien and Toms (2010) developed a construct for evaluating user engagement and novelty based on the state-trait curiosity model that considers external and internal stimuli, as well as individual differences. On the other side, Attfield et al. (2011) surveyed different aspects of user engagement and, although they consider it very context dependent, they discuss how the user's expertise shapes the engagement by enabling more control over the richness or potential of the system. Investigating whether these results apply in the same way to recommender systems by providing user controllability motivates this research question.

RQ4. Do user characteristics affect the influence of controllability on the user's experience in a recommender system?

The influence of user characteristics on the user's experience has been already studied in music, e-commerce and energy-saving recommenders. The motivation for this research question stems from investigating if the same user characteristics previously studied affect the user experience in the context of conference's talk recommendation by interacting with the role of controllability and also by considering a more comprehensive evaluation. For one thing, controllability and inspectability in a hybrid recommender are implemented in a different way in SetFusion (Venn diagram and sliders instead of only sliders or user-controlled sorting of lists) compared to previous implementations (Bostandjiev et al., 2012; Knijnenburg et al., 2011); hence the results on user experience may not necessarily hold.

4.3. Evaluation Measures and User Characteristics

4.3.1. Evaluating User Engagement

To measure the effect of controllability on the user's engagement with the system, we used subjective and objective measures based on the guidelines presented in O'Brien and Toms (2010) and Attfield et al. (2011).

Subjective Measures. These measures were captured by questionnaires. Many of these factors or constructs (groups of questions) are suitable for evaluating the recommender user experience in general as well. O'Brien and Toms (2010) identified six final constructs that play a role in evaluating user engagement; we considered the four that were most relevant for our study: (a) focused attention; (b) perceived usability; (c) endurability; and (d) novelty.

Objective Measures. These metrics were measured by direct observation of user actions rather than by interpretation of their perception of opinion, such as a survey. The individual behavioral measured the recommender interface: (a) number of talks explored; (b) number of talks bookmarked; (c) number of clicks; and (d) amount of time.

4.3.2. Evaluating User Experience

Evaluating the users' experience in a recommender system in a holistic way is not new, but only recently has the research community proposed frameworks to guide evaluation and make results more easily comparable. After the initial work of McNee et al. (2006a) introducing the Human-Recommender Interaction model (HRI), two more elaborated user-centric evaluation frameworks for recommender systems have been proposed. ResQue, introduced by Pu et al. (2011), and one introduced by Knijnenburg et al. (2012). Building on the aforementioned studies, these are the perceived systems qualities that we evaluated through surveys in our research:

- Related to perceived system qualities: explanation, interaction adequacy, recommendation Accuracy, recommendation diversity, information sufficiency, interface adequacy.
- Related to user beliefs: transparency, control, perceived usefulness, perceived ease of use.
- Related to user attitudes: trust and confidence, overall satisfaction
- Related to behavioral intention: user intention.

Objective metrics

In addition to the metrics described in the aforementioned user frameworks, we used traditional measures of evaluation of recommender systems:

- *Average rating*: we compare the conditions by calculating the mean over the average rating of each user under a particular condition.
- *Precision@k*: this metric allow us to measure the accuracy of a list of *k* recommendations (Manning, Raghavan, & Schtze, 2008).
- *MAP*: Mean Average Precision (Manning et al., 2008) is a metric that calculates the mean over the *average precision* of several lists. The average precision of one list is calculated by averaging the precision at several cut points, usually the recall points (the positions of the list where the element found is relevant).
- *MRR*: stands for Mean Reciprocal Rank. It is calculated as the inverse of the ranking position of the first relevant element to be found on a list (Manning et al., 2008).
- *nDCG*: stands for normalized Discounted Cumulative Gain (Manning et al., 2008). This metric allows us to tell how well the recommender system ranks a list of recommendations. If the ranking is perfect, the relevant recommendations will be at the top of the list and the non-relevant at the bottom, resulting in a nDCG = 1.

4.3.3. User characteristics

Based on previous studies, we collected information about the following personal and situational characteristics to investigate research questions three and four:

- User expertise in her own domain: Is the user knowledgeable in her own domain?
- Familiarity with iConference: How familiar is the user with the user community of iConference?

- User experience with the system: Has the user used Conference Navigator 3 before?
- Trusting Propensity: Does the user have an inherent propensity to trust in people or systems?
- User experience with recommendation systems: Does the user have some previous experience or knowledge about recommender systems?

4.4. Participants

Subjects were recruited by e-mail and by ads posted at the School of Information Sciences and the School of Computer Science at the University of Pittsburgh, and also at the Heinz College at Carnegie Mellon University. Three promotional e-mails were also sent to mailing lists of graduate students of Library Science, Information Sciences and Telecommunications, and the Intelligent Systems Program at the University of Pittsburgh. The main requirement was that they should have a clear interest in reading research articles, most of them had already earned a PhD or were pursuing a PhD, in areas related to the iConference (e.g., social media, social computing, social networks, IT policy, etc.). Each subject received an incentive of $12/hour for participating in the user study. In order to prevent a judgment biased in favor of papers already known from past conferences, participants should have attended none or at most one *iConference* in the last 3 years (*iConference* 2011, 2012, and 2013). In the case of users that had already attended one *iConference* (for instance, *iConference* 2012) the attended conference was used as the *seed conference* on the *Pref* step to identify user interests and to generate recommendations for the other two conferences (in this case, *iConference* 2011 and 2013). The subjects were assigned to an order in which controllable and baseline interfaces were presented to one of several conference sequences ensuring an appropriate balance among the conditions. Table 1 presents a summary of the number of subjects under each condition:

Seed iConference	Controllable interface iConference	Baseline interface iConference	Number of subjects at sequence order	
			Controllable => Baseline	Baseline => Controllable
2011	2012	2013	3	3
2011	2013	2012	3	3
2012	2011	2013	4	4
2012	2013	2011	3	3
2013	2011	2012	4	4
2013	2012	2011	3	3

Table 1. Distribution of the 40 participants over the different conditions.

4.5. Instrumentation Details

The SetFusion and the baseline interface used the same recommendation approaches in the background to suggest articles to CN3 users. Details of these recommenders and how they were combined in each condition (baseline and user-controllable) are explained in this section.

4.5.1. Recommendation Approaches

User control in SetFusion was implemented by letting users combine different recommendation algorithms. Given these considerations, we used information crawled from the ACM library to fast start

(Brusilovsky, Parra, Sahebi, & Wongchokprasitti, 2010) users' and items' profiles, which allowed us to produce recommendations for the popularity, content-based and collaborative filtering methods.

Bookmarking Popularity

In this study we used three conferences hosted by Conference Navigator (CN3): iConference 2011, 2012, and 2013. The system stored the information about how many people bookmarked each paper in each conference. We used this popularity as one of the recommendation algorithms. It is not personalized, but it leverages the social wisdom of actual conference attendees.

Content-Based Algorithm

Considering that the subject has provided feedback (bookmarks) to papers in a related conference, we used the title and abstract of those papers as a "user model" and found similar papers in the current conference to produce the recommendations. The user model consisted of a vector of terms made from the titles and abstracts of the papers that the user had chosen, where the weights of the vector were calculated by TF*IDF (term frequency * inverse document frequency), as explained in Manning, et al. (2008). To find relevant documents, the matching was performed using the Lucene[2] function MoreLikeThis[3], which performs a cosine similarity matching between the user profile, represented as a vector or terms, and the talks in the conference index, returning a list of the most related documents.

Author-based Popularity

How many times the authors have been cited can be considered another source of article relevance. This means that popularity was understood as *expected impact popularity*. To calculate the popularity of papers' authors based on how frequently they were cited in the past, we used a dataset crawled from the ACM Digital Library. Starting from there, the procedure to obtain the popularity of each paper:

a) List the papers for each conference in CN3
b) Obtain the author names from the papers found in (a)
c) Match the author names with the author names in the ACM database
d) For each author matched in the ACM DB, obtain the number of references
e) Calculate the popularity of each paper found in (a) by aggregating the number of references of each of its authors as found in (d). By aggregation, we mean a function that gives a relevance score to a paper based on the maximum "number of references" among the authors of that paper.

4.5.2. Fusing the Recommendation Approaches

Each recommendation method explained in the previous section returned a set of papers ranked by a relevance score. It was our goal to combine these recommendation sets, but the relevance scores returned by the three methods were in different ranges. In one case the score goes from 0 to 1, in another case from 1 to 10, and the normalized scores distributions are not similar. Then, the final score of each recommended item was based on its rank in the recommendation list of each method, and with more importance given to items recommended by more than one method. The fusion was performed in such a way that the score of a recommended item $src(rec_i)$ was given by:

[2] http://lucene.apache.org/

[3] https://wiki.apache.org/solr/MoreLikeThisHandler

$$src(rec_i) = \left[\sum_{m_m \in M} \frac{1}{rank_{rec_i, m_j}} \times W_{m_j} \right] \times \left| M_{rec_i} \right|$$

Equation 1. Score function for the hybrid recommender system.

This is a similar formula for a cross-sourced hybrid recommender as used in Bostandjev et al (2012), where M is the set of all methods available to fuse –in our work: bookmarking popularity, content-based, and author-based popularity-, $rank_{reci,mj}$ is the rank –position in the list- of recommended item rec_i using the method m_j, W_{mj} corresponds to the weight given by the user to the method m_j using the controllable interface, and $|M_{reci}|$ represents the number of methods by which item rec_i was recommended. In the case of the baseline interface where users could not control these weights, and after conducting a 10-fold cross validation with the results obtained in Parra et al. (2013), we set the weights to the non-controllable recommender to 0.4 for bookmarking popularity, 0.6 for context-based, and 0.4 for author-based popularity.

4.6. Study Procedure Details

Since all elements of the study design have been introduced, we now present the study procedure in detail. The workflow was as follows:

a) Participants signed an informed consent of the benefits and risks of the study, as specified in the IRB.
b) Participants completed a pre-questionnaire that told us demographic information, progress and experience in their graduate program, familiarity with the iConference, familiarity with Conference Navigator, trusting propensity and familiarity with recommender systems.
c) **Task 1 (Bookmarking task):** The participant was given a hypothetical situation where she was attending the iConference and one week before the conference takes place her advisor asked her to identify the most relevant papers for her as well as for colleagues in the same laboratory. Participants were told that the main purpose of including "colleagues" was to increase the number of relevant papers, because the iConference is very diverse and there is a chance that too few papers were relevant only for her.
    1. Subtask 1 (obtain user preferences): By scanning each paper in the proceedings of one iConference, the subject must identify the papers relevant for her and for some colleagues of her choice, without being limited by time or in the number of papers to be judged as relevant. After this step is finished, in a different screen, the user must state for each paper selected if it was judged as relevant only for her, for her and her colleagues, or only for her colleagues.
    2. Subtask 2: Using the controllable or the baseline recommender interface, the subject must find at least 15 papers relevant for her, for her colleagues or for both. After selecting the 15 papers, the subject indicates in a separate screen for whom each bookmarked paper is relevant (for herself, colleague or both). After finishing this step, the user must answer a post-session survey regarding the interface (controllable or not) she was assigned in the first task. This survey is in the Appendix section of this article.
    3. Subtask 3: The subtask 2 along with the post-session survey is repeated for the other recommender interface –controllable or not-controllable- depending on the condition the user had been assigned to.
d) **Task 2 (Ratings)**: Rate all critical papers in all three conferences used in the study on a scale from one to five, where one means not relevant at all and five means strongly relevant (Figure 8). It

includes all papers of the seeding conference and all papers shown in any of the recommended talk lists in each of the two explored interfaces. The papers are sorted randomly and the icon beside the title indicates if the paper was bookmarked in the previous task.

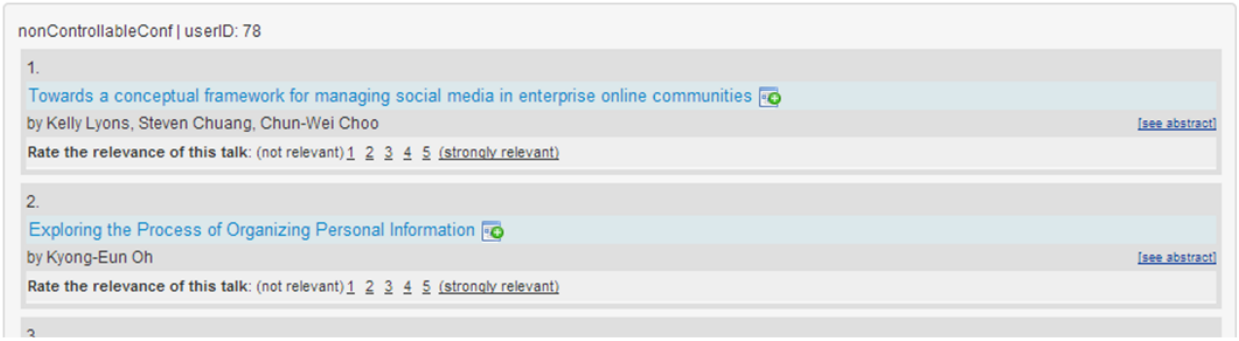


Figure 8. Screenshot of the rating interface.

e) Post-questionnaire: Obtain participants' perceptions about both recommendation interfaces, asking them to judge which one was preferred over the other, which they would advise to permanently implement in Conference Navigator, and which one required more effort in order to finish the requested task. Subjects were also asked to elaborate freely by writing or by talking (in this second case the answer was recorded) why they preferred one interface over the other.

5. Results of Behavioral Analysis

To acquire an overall understanding of how users utilized the controllable interface compared to a traditional static set of recommendations, we started by analyzing how many users tried each feature available on both interfaces, then whether people used the controllable widgets to bookmark conference articles, and finally if the rating of the articles bookmarked was influenced by the filtering methods used by the subjects. The user actions logged in each interface with its respective description and associated visual component are presented in Table 2.

Action	**Description**	**Visual Widget**	**Controllable Interface**	**Baseline Interface**
clickRetrieveList	Retrieve initial list of recommendations	Recommender interface	X	X
Scheduling	Bookmark a talk	Article	X	X
Unscheduling	Remove bookmark	Article	X	X
seeMore	Expand article list	Recommender List	X	X
clickOpenAbstract	Open abstract of talk	Article	X	X
clickCloseAbstract	Close abstract of talk	Article	X	X
changeSlider[N]	Change weight of method N	Slider Widget	X	
clickUpdateList	Update recommendation list	Slider Widget	X	
hoverMethod[N]Explain	Dialog Explains method N	Slider Widget	X	
hoverCircle[N]	Mouse over circle (talk) on the	Venn diagram	X	

	subarea (method) N			
clickEllipse[N]	Click on Venn diagram to filter list by method N	Venn diagram, list of talks	X	

Table 2. List of actions tracked in the recommender interfaces (Controllable and Baseline).

5.1. How many subjects use each feature on the recommender interfaces?

Figure 9 shows the number of subjects that used each action available on both interfaces, with the x-axis showing the number of users and the y-axis the action names. The red color is used for counts of subjects in the baseline interface and green bars for the actions of the controllable interface. The blue and green boxes on the y-axis are used to indicate the related actions. The total number of subjects in the study was 40, and, as a within subjects study, all subjects experienced both conditions (interfaces): see that 40 people performed the actions *clickRetrieveList* (the action that loads the lists of recommendations) and *scheduling* (the action needed to finish bookmarking talks).

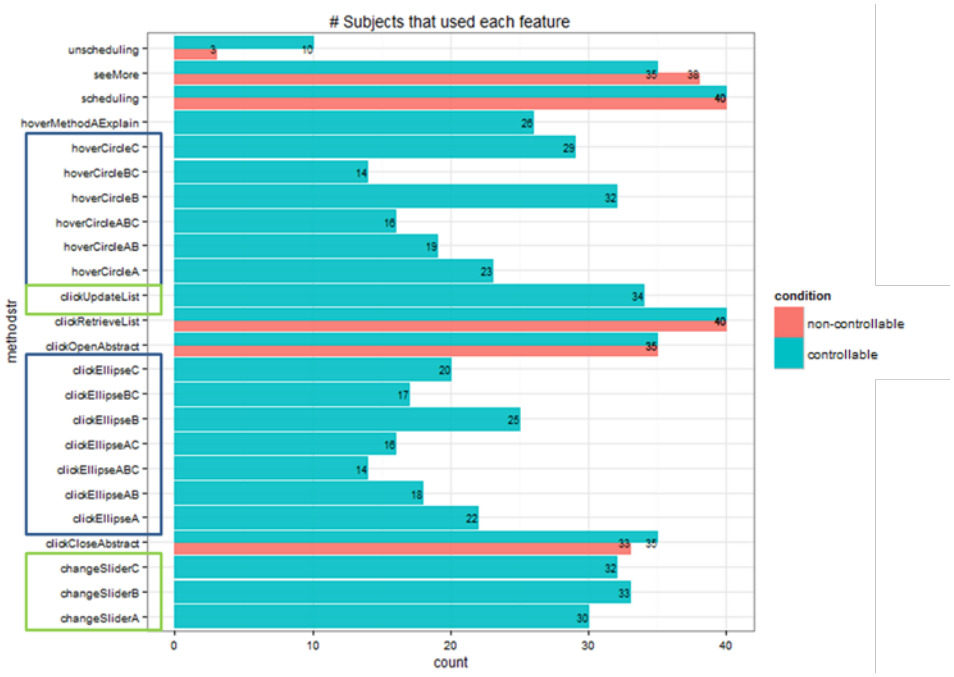


Figure 9. Plot of the amount of subjects that used each action available in the recommender interfaces.

95% of the subjects in the baseline interface expanded the list of recommendations with the action *SeeMore*, compared to 87.5% of subjects on the controllable interface. Also, 87.5% of people under both conditions *opened the abstracts* during the bookmarking task and the same percentage *closed the abstract*s in the controllable interface, compared to 82.5% of people who *closed the abstract* in the baseline interface. Under the controllable interface, more people used the sliders than the Venn diagram features. Figure 9 shows that among the actions available in the sliders, at least 75% people used some *changeSlider[N] action* –to set the importance of a recommendation method. With respect to the Venn diagram, up to 80% of the subjects used hovering (positioning the mouse over circles on the content-based recommender to display its title on a floating dialog) and up to 62.5% use filtering. Filtering with Venn diagram overlapping areas was used less than full area filtering; yet even *clickEllipseABC* (clicking on the intersection area of the three ellipses that filters the papers in the list recommended by

the three methods) was used by 35% of subjects. Limited use of overlaps is not strange since these areas contain very few, frequently no talks, though we later see that papers in this intersections have the largest user rating in average.

5.2. Do people use the sliders and Venn diagram to bookmark talks?

While not all subjects used the provided controls, our data shows that for those who used it, sliders and filters provided an efficient tool for finding good talks. Under the controllable interface, the subjects had the chance of bookmarking papers directly from the recommended list, dismissing the visual widgets. They could also "generate" sub lists of recommendations by filtering (clicking) on different areas of the Venn diagram, or by changing the weight with the sliders. There were a total of 616 articles bookmarked (625 bookmarking actions, but people could also *unbookmark*) in the controllable recommender interface, and as Table 3 shows, the vast majority of these bookmarks were submitted either after using sliders to re-rank the results or after using the Venn diagram to filter it. Using a one-sample test of proportions we found that the fraction of bookmarks performed after using each tool, re-ranking or filtering, is significantly larger than the 13.47% of talks bookmarked without any tool, χ^2=654.55, $p <$ 0.001. Moreover, significantly more than half of the talks were bookmarked after using the slider control. This proportion (58.44%) is significantly larger than half of the talks (50%), after performing a one-sample test of proportions, χ^2=8.5, p = 0.003.

Without Filters	Using sliders	Using Venn diagram
83 (13.47%*^)	360 (58.44%^)	173 (28.08%*)

Table 3. Distribution of bookmarks by filtering method in the controllable interface. Tests of proportions show that the percentage of bookmarks without filters is significantly smaller than bookmarks with sliders actions (^p < 0.001) and also smaller than percentage of bookmarks with Venn diagram actions (*p < 0.001)

5.3. Are fusions of recommendations useful for finding good talks?

As shown in Figure 9, subjects used filtering by overlapping areas more rarely than filtering by whole areas. But was it really useful to click on overlaps? Did it bring better talks to the surface? Figure 10 shows a plot of average ratings of talks suggested by different recommendation methods (or combinations of them) separated by condition. The data indicates that it was not the single method that was the most productive (in bringing best talks), but the combinations of AB and especially ABC. Note that it was not the presence of these talks on the overlapped areas on the Venn diagram that caused the subjects to rate these talks so highly; for subjects in the non-controllable condition who had no way to see which talk was recommended by which method these talks also have the highest rating. Yet, the presence of the Venn visualization apparently helped the subjects to recognize the overlapping areas as the most productive, producing a visible boost of ratings in the overlapped areas in the SetFusion condition. The statistical analysis confirms both effects. Within the baseline interface, the average user rating of talks bookmarked with fusion of methods ABC (M = 2.96, S.E. = 0.29), was significantly higher than method C (M = 2.29, S.E. = 0.21, p = 0.005), but not higher than methods A (M = 2.36, S.E. = 0.2, p = 0.072) and B (M = 2.57, S.E. = 0.22, p=0.331), using a related-samples Wilcoxon Signed rank test. On the other side, under the controllable interface, the average user rating of methods ABC (M = 3.28, S.E. = 0.27) was significantly larger than methods A (M = 2.38, S.E. = 0.2, p=0.002), and C (M = 2.17, S.E. = 0.2, p<0.001), but not larger than B (M = 2.51, S.E. = 0.21, p = 0.054) using a related-samples Wilcoxon Signed rank test.

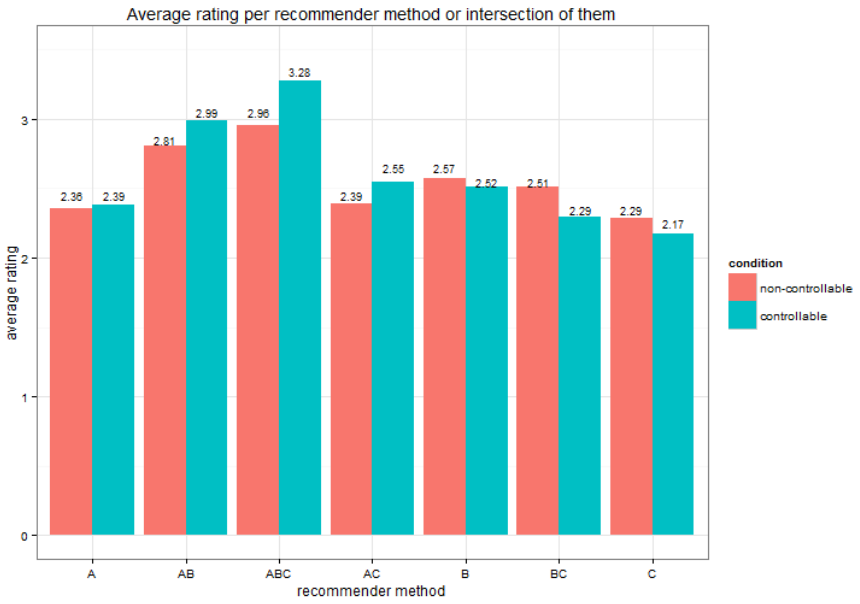


Figure 10. Average rating per recommender method (or overlaps of them) under the non-controllable and controllable interfaces. *A* is popularity based on bookmarks, *B* is the content-based recommender, and *C* is popularity based on authors' citations. In overlaps, *AB* means papers recommended by both methods A and B.

Summary of Behavioral Analysis. The results of this section provide evidence that subjects actively used the features available for controllability. There is also evidence that the controllability filters (sliders and Venn diagram) were useful to find relevant papers among the recommended items, implying that user control and transparency were effective tools to find relevant items. We also found that the items located in the overlapping areas of the Venn diagram tend to be of better relevance as indicated by user ratings. Moreover, there is evidence that the presence of the Venn diagram amplified the value of these talks for the users.

6. Quantitative Analysis on Objective and Subjective Metrics

We designed and conducted our study with the assumption that the controllable interface would show a better engagement and user experience than the baseline interface. To find a possible value of SetFusion, we first compared its impact on the most evident parameters, number of bookmarks and average rating. In our within-subjects study with 40 participants, the total number of bookmarking actions in the baseline interface was 638 (15.95 bookmarks per user) with an average rating of 2.48±0.089, while the controllable interface had 625 bookmarks in total (15.63 bookmarks per user) with an average rating of 2.46±0.076. On that general level, there were no significant differences between interfaces. However, when the order of interfaces was considered, we found users that were shown the controllable interface after the baseline spent significantly more time with the recommender. In Figure 11, the time spent with the controllable interface (green line) is almost invariable whether presented first or second, yet the baseline interface (blue line) shows a significant drop in time when presented after the controllable interface. This behavior can be interpreted either as a good or bad sign of engagement with the user-controllable interface: users might have felt engaged with the features and spent more time exploring them, or rather they felt confused with the interface, decreasing their performance.

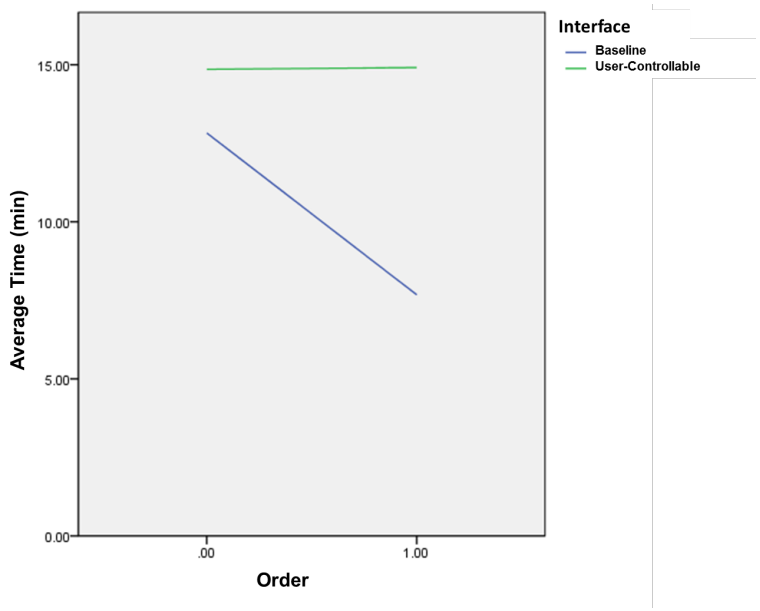


Figure 11. Effect of the treatment interface (User-Controllable, green line) and the order in which it was presented (x-axis) on the amount of time (y-axis) in minutes that users spent in the bookmarking task.

The previous results indicated an influence of interface and order on user performance, but deeper analysis is necessary to find which factors motivated this difference and what it really meant in terms of user perception. In order to complete such an analysis, we first built features (constructs) from user characteristics by conducting a factor analysis over the answers on the pre-study questionnaire. After identifying these constructs that would be used as factors in regression models, we conducted three sets of regressions. We call a *set* of regressions to a group of regression models built using the same random and fixed factors, but different independent variables. We connected the three sets of regression through a block model, seen in Figure 12, and we describe it as follows:

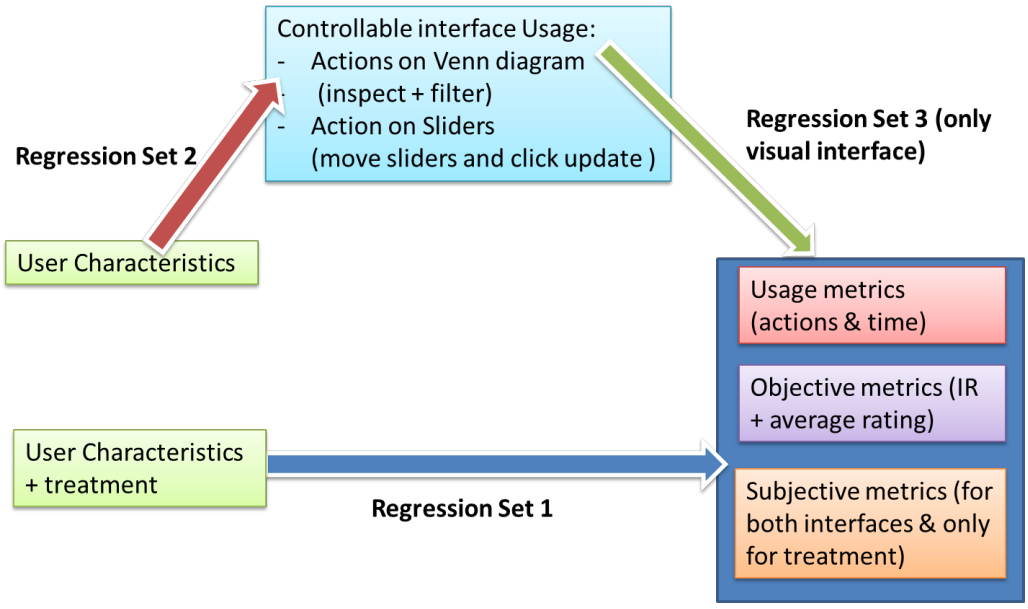


Figure 12. Block model summarizes the statistical analysis conducted in the user study

Regression Set 1: In this set we directly tested the effect of the interface and we also controlled for user characteristics over several performance metrics. Our fixed factors were user characteristics, the interface, the order in which the interface was presented, and the conference itself; the user was modeled as a random factor. The outcome variables were usage metrics, objective metrics and subjective metrics.

Regression Set 2: In this set we focused on the SetFusion interface. We built a regression model to predict the number of certain actions performed by the participants in the SetFusion interface as a function of user characteristics. The fixed factors were then user characteristics and the outcome variables were the number of actions either on the sliders widget or on the Venn diagram widget.

Regression Set 3: After predicting which user characteristics might explain the use of certain actions on the controllable interface, we then used these actions (behavioral metrics) as predictors of objective and subjective measures. As in Regression Set 2, these regression models focused on the SetFusion controllable interface.

The rest of this section is structured as follows: 6.1 describes the factor analysis to derive constructs of user characteristics from the pre-study survey. The next three sections describe the results of the three regression analyses respectively: 6.2 Regression Set 1, 6.3 Regression Set 2, and 6.4 the Regression Set 3.

6.1. Factor Analysis: User characteristics extracted from Pre-Study Survey

Study participants answered 19 questions (details in the Appendix), the first four about demographics (occupation/program, gender, native English speaker, age), with the other 15 intended to asses five characteristics: expertise in her own research domain, engagement with the iConference community, familiarity with the system (Conference Navigator), trusting propensity, and familiarity with recommender systems. All the previous characteristics have shown some effect on the user experience in previous studies, e.g., (Knijnenburg, Bostandjiev, et al., 2012; Knijnenburg et al., 2011; Pu et al., 2011).

Question\Latent Factor	**Trusting Propensity**	**Research Expertise**	**Familiarity with Recommenders**	**Trust in Recommenders**
In general, people really do care about the well-being of others	0.889	0.255	-0.208	0.311
The typical person is sincerely concerned about the problems of others	0.766	-0.127	-0.188	0.128
Most of the time, people care enough to try to be helpful, rather than just looking out for themselves	0.906	-0.172	0.224	-0.307
If you are pursuing a PhD degree, which stages have you completed in your program of study?	0	0.993	0	0
How many conference or journal papers have you published in your area of research?	0	0.663	0	0.180
I am familiar with online recommender systems	0	0.284	0.809	0
I know of one or more methods used to produce recommendations in a system	-0.152	-0.108	0.864	0

I have occasionally followed the advice of a recommender system (such as a recommended book in Amazon.com or a recommended video in YouTube)	0	0.112	0	0.986

Table 4. Questions and their loadings on the latent factors resultant of the EFA. Maximum likelihood estimation used as factor extraction method and varimax rotation.

The exploratory factor analysis showed that some questions did not load well on any of the factor models fitted, so they were removed. The final factor analysis model had four factors, and their respective loadings can be seen in Table 4. In order to create a unique composite score for each latent factor, the answers of the related questions were standardized and then averaged, with the exception of the factor "Trust in Recommenders", which comprises only one question. We also kept in the list of final user characteristics whether the user had previous experience with CN, since the single answer yes or no to this question suffices to measure that property. This is not the case with trusting propensity, a more complex factor that requires multiple questions to be assessed accurately. Finally, the variables considered in this study to control for the effects of user characteristics are: (a) Occupation (PhD student in different areas, Postdoc, researcher, lecturer), (b) Age (continuous variable), (c) Gender (male/female), (d) Research Expertise: construct of two standardized variables, (d) Trusting Propensity: construct of three standardized variables, (e) Experience with Recommender Systems: construct of two standardized variables, (f) Trust in Recommender Systems: standardized variable of one question, and (g) Previous use of Conference Navigator (yes/no).

6.2. Regression Set 1: The Effect of User Characteristics and Interface on Several Dimensions

In this regression set we analyzed the effect of the interface (treatment) and user characteristics over three types of outcome or dependent variables, as shown in Figure 13. For the sake of space, we only reported those factors (independent variables) that had a significant effect over the metrics studied as outcome: usage metrics (amount of talks explored and time), objective metrics (MAP, MRR, Precision@5), and subjective metrics (the metrics represented by questions in the post-session surveys)

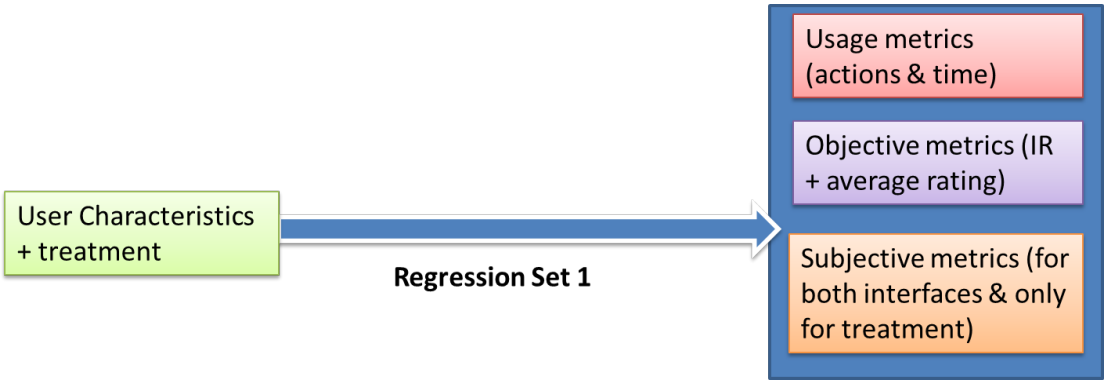


Figure 13. Block model summarizes first set of regressions.

Regressions on Usage Metrics

The two usage metrics considered in this analysis, shown in Table 5, are comparable between conditions: (a) Talks explored using mouse actions (not considering Venn diagram hover actions), and (b) Time Spent. Considering the distribution of the outcome variables, a negative binomial regression was

conducted to understand the effect of the interface and user characteristics on the numbers of talks explored, whereas a gamma regression was conducted to study the effect of the aforementioned variables on the time spent in the bookmarking task.

D.V.	I.V.	β	exp(β)	*p-value*
Talks Explored	Experience with recsys	-0.1	0.91	0.033
	Experience with CN	0.31	1.36	0.044
Time Spent	Order	-0.54	0.58	< 0.001
	Interface*Order	0.6	1.82	0.002

Table 5. Significant effects of the regressions on usage metrics

In the case of the metric *talks explored,* measured through actions available in both interfaces (*bookmark talk*, *open abstract*, *close abstract*), the treatment (the controllable recommender interface) did not have a significant effect, while user characteristics did. Users with previous experience and knowledge of recommender systems explored significantly fewer talks, $p = 0.033$. A one unit increase in standardized experience decreases the number of explored talks in a factor of 0.91, i.e., a 9% decrease. This is understandable since these users can rely less on the content of the paper and more on the explanations (the method or methods used to recommend) to judge the relevancy of a talk. On the other hand, being familiar with the conference system CN3 increases the number of actions 36% compared to those who are unfamiliar, $p = 0.044$. Given that the list of items is presented in a similar design to other pages in the system (e.g., proceedings, top items, etc.), users familiar with the system will be more likely to try features with similar functionality to explore talks, since they know what to expect from them, decreasing their cognitive strain in exploring the items (Albers, 1997). Regarding time spent, we see here the effect described at the beginning of this section regarding the order in which the interface was used to perform the task. If we set the time spent on the baseline interface when it was presented first (before the controllable) as the reference point, we observe a decrease in time spent in a factor of 0.58, i.e., a 42% decrease when the baseline interface is presented second. On the other hand, if the interface presented the second is the user-controllable one, there is actually an increase of 82% in the time spent on the interface.

Regressions on IR metrics

Linear mixed-model regressions were conducted to understand the effect of the treatment (recommender interface) and user characteristics on average user rating, MRR, MAP, nDCG, and precision@n (n=3 and n=5). The significant effects found on these regressions are shown in Table 6. User average rating had no significant effects, and the significant effects of nDCG and precision@3 were driven solely by one outlier, so they were removed.

D.V.	I.V.	β	*p-value*
MAP	Interface	0.08	0.016
MRR	Order	0.21	0.021
Precision@5	Gender (male)	0.45	0.005

Table 6. Significant effects of the regressions on accuracy and IR metrics

The controllable interface produced a significant increase of 0.08 in the MAP compared to the baseline. While the MAP increase could be considered as evidence that the controllable interface delivered better

performance, it is important to remember that our study featured a condition that is rare in classic IR: a single list in the non-controllable interface is compared to a set of dynamic lists in the controllable recommender. As a part of further work, it would be interesting to explore other metrics such as recently introduced IR metrics for interactive user sessions. MRR is a metric that attempts to measure how good the ranking algorithm is at locating the first relevant item at the very top. In our study, MRR is 0.21 units, significantly higher when users bookmark with the second interface, independent of being the baseline or the controllable. This result can imply that users have become more familiar with the interface and task and they were able to find relevant papers at higher ranks. The role of gender in increasing the precision at cut point 5 in 0.12 units is not easy to interpret, but some hints were found when analyzing the results of the post-session survey with the perception of several male participants about the role of the Venn diagram.

Regressions on Subjective Metrics

In this set of regressions, the outcome variables were the concepts associated with each question in the post-session surveys (see details in Appendix, section 11.2). Table 7 shows four columns, the first one with a shortcut of the statement evaluated and the second column showing significant effects explaining the variability of the metric. The first ten metrics (from UNDERSTOOD to RECSYS_NO_NEED) compare both interfaces, and metrics 11-17 (from C_FEEL_CONTROL to C_VENN_TRUST) refer only to features in the controllable interface. Each shortcut and the associated statement are explained in *Appendix*. This set of regressions was conducted using a linear mixed-model analysis with a Gaussian identity link. The data shows some interesting impacts of both the interface and user features on user opinion.

Metric	Question	Significant effects	β	p-value
UNDERSTOOD	I understood why the talks were recommended to me.	Use of CN	0.94	<0.001
		Interface	1.1	<0.001
		Native Speaker	0.68	0.003
DIVERSE	The items recommended were diverse.	Experience in research domain	0.15	0.04
INTERFACE_EASY	I became familiar with the recommender interface very quickly	Order (second)	-0.53	0.014
		Interface(treatment) *Order(second)	0.97	0.012
LOST_TRACK_TIME	I lost track of time while I was using the recommender interface	Experience in research domain	-0.26	0.039
OVERALL_SATISFIED	Overall, I am satisfied with the recommender interface	Order (second)	-0.97	<0.001
		Interface(treatment) *Order(second)	1.09,	0.025
CONFIDENT_MISS	The recommender made me more confident that I didn't miss relevant talks	Order (second)	-0.92	<0.001
		Interface(treatment) *Order(second)	1.19	0.012
USE_AGAIN	I would use this recommender system again for another conference in the future	Order (second)	-0.98	<0.001
		Interface(treatment)	1.1	0.029

		*Order(second)		
SUGGEST_COLLEAGUES	I would suggest my colleagues to use this recommender system when they attend a conference in the future	Order (second)	-1.22	< 0.001
		Interface(treatment) *order(second)	1.54	0.004
RECSYS_NO_NEED	I do not think that a social conference support system - like Conference Navigator- needs Talk Recommendation functionality	Experience with recsys	-0.24	0.046
C_FEEL_CONTROL	I felt in control of combining different recommendation methods by using the sliders.	Order	0.79	0.014
C_VENN_UNDERSTAND	I think the Venn diagram visualization helped me to understand why a talk was recommended.	Gender	-0.74	0.018
C_VENN_USE	I think the Venn diagram visualization was useful to identify talks recommended by a specific recommendation method or by a combination of recommendation methods.	Trusting propensity	0.13	0.03
C_VENN_TRUST	The ability to use the Venn diagram to examine the talks recommended increases my trust in the list of recommended talks.	Trusting propensity	0.19	0.004

Table 7. Significant factors on subjective metrics collected in post-session survey.

- Using SetFusion increases the agreement with understanding why talks were recommended in 1.1, $p < 0.001$, keeping the other variables constant. Being a native speaker ($\beta = 0.68$, $p = 0.003$), and having previous experience with CN ($\beta = 0.94$, $p < 0.001$) also positively influenced this variable.
- Using SetFusion *after the baseline* interface improved the user perception on 5 out of 10 metrics: getting familiar with the interface quickly ($\beta = 0.97$, $p = 0.012$), being satisfied overall with the interface ($\beta = 1.09$, $p = 0.025$), being confident of not missing relevant talks ($\beta = 1.19$, $p = 0.012$) and also the user's intention of using it again ($\beta = 1.1$, $p = 0.029$) and suggest the interface to colleagues ($\beta = 1.54$, $p = 0.004$)
- Having more experience in the research domain resulted in a *better* perception of the diversity of the talks ($\beta = 0.15$, $p = 0.04$), but also a *lower* perception of engagement - less feeling that the task makes them lose the track of time ($\beta = -0.26$, $p = 0.039$).
- The experience with recommender systems significantly decreases the perception that CN *does not need* a recommender system ($\beta = -0.24$, $p = 0.046$), i.e., users more appreciate the presence of the recommender in CN.
- The subjects' trusting propensity caused a positive impact on perceiving Venn diagram as useful to identify talks recommended by different methods ($\beta = 0.13$, $p = 0.03$), also increasing the trust in the recommendations given the use of the Venn diagram ($\beta = 0.19$, $p = 0.004$).
- Women perceived that the Venn diagram helped them understand why the talks were recommended significantly more than men did ($\beta = -0.74$, $p = 0.018$).

6.3. Regression Set 2: The Effect of User Characteristics on User Behavior in the Controllable Widgets

In this regression set we focused on the actions only available in the user-controllable interface, specifically, on the Venn diagrams and the sliders. Our expectation was that some user characteristics could affect their use of control and inspectability features. By performing negative binomial regression

analyses on the number of actions on the Venn diagram and another regression on the number of actions on the sliders widget, we found significant effects of some user characteristics, as shown in Table 8:

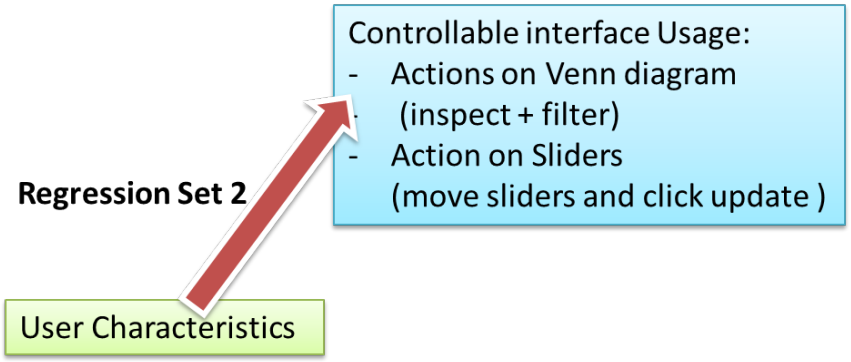


Figure 14. Block model of the effects and dependent variables studied in the second set of regressions.

D.V.	I.V.	β	exp(β)	*p-value*
Venn actions	Native speaker	-0.69	0.50	0.038
Sliders actions	Native speaker	-0.67	0.51	0.032
	Trust in Recsys	0.37	1.44	0.013
	CN Use	0.6	1.82	0.032

Table 8. Significant effects of the regressions on actions over the visual widgets.

- Native speakers performed significantly less actions than non-native speakers with both the Venn diagram (50% decrease) and sliders (49% decrease). We can speculate that the talk selection was harder for non-native speakers who need all support that the system can provide in this task. On the other hand, since the list of recommendations was not large (from 30 to 60 papers), the task was apparently not sufficiently hard for native speakers who could rely more on fast scanning of the talk information (title, author, paper type) at the recommendation list.
- Higher trust in recommender systems increases the number of actions on the sliders by 44%. While the talk selection task might not be as hard as to require using the sliders, having trust in the recommendations produced by the algorithms motivate people to invest more time in practicing and using the sliders.
- Previous use of Conference Navigator increased the use of the sliders by 82% compared to those without previous experience using sliders. We can speculate that having previous experience with the system decreased the overall cognitive load encouraging participants to try new actions (Albers, 1997).

6.4. Regression Set 3: Regression on Evaluation Measures Controlling for Actions on the Interactive Interface

To understand whether specific actions with the controllable interface had an influence on the user experience, the logarithm of the number of actions on the Venn diagram and on the sliders widget were

used as predictors of objective and subjective metrics, but this time considering only the controllable interface. We did not consider usage metrics such as clicks on the recommended list since we did not have a reliable assumption on how one type of actions could explain the variability of the other type. Figure 15 summarizes the analysis in a block model. The decision to use the logarithm instead of the raw number of actions is based on previous literature (Hu, Koren, & Volinsky, 2008; Marujo, Bugalho, Neto, Gershman, & Carbonell, 2013) that showed a better performance in prediction tasks. The regressions performed were controlled for both variables (Venn actions and sliders actions) as predictors since the correlation between them is not significantly different than zero, rho = 0.17, $p = 0.3$

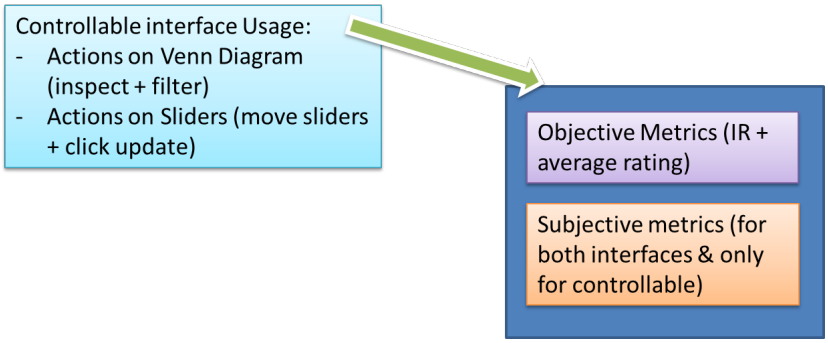


Figure 15. Block model of the regression on evaluation measures controlling for actions on the controllable recommender interface. Dependent variables are objective and subjective metrics, usage metrics are not considered.

6.4.1. Regressions on IR metrics

Linear multiple regression was used to study the effect of the number of actions on the Venn diagrams and on the sliders upon the ranking metrics MAP, MRR, nDCG and the accuracy metrics precision, precision@n (n=3 and n=5), and average user rating on the controllable interface.

Neither the amount of log-actions on the sliders nor on the Venn diagram had a significant effect when we considered only the controllable interface (see details in the Appendix). Although we did not find significant factors, we must emphasize that an extended analysis using new metrics designed to evaluate interactive sessions (rather than static lists) could eventually reveal some effect. All the IR metrics studied were designed to evaluate a single static list of relevant items, and recently some works have proposed multiple-query session IR metrics, like sDCG (K. Järvelin, S. Price, L. L. Delcambre, & M. Nielsen, 2008).

6.4.2. Regressions on Subjective metrics

Our final set of regression assumed that the actions performed over the control widgets, sliders and Venn diagram, could have a direct effect on the perception that participants have over the system, such as their understanding, their perception of control or their general satisfaction. When performing the regression on the survey metrics controlling for log-actions on sliders and on the Venn diagram, both show a significant effect on two statements referring to the Venn diagram, as shown in Table 9. The results show a competition between the use of the Venn diagram and the use of sliders on these two statements.

Metric	Significant effects	β	p-value
I think the Venn diagram visualization helped me to understand why a talk was recommended.	Venn_actions,	0.17	0.031
	Sliders actions	-0.27	0.012
I think the Venn diagram visualization was useful to identify talks recommended by a specific recommendation method or by a combination of recommendation methods.	Venn_actions,	0.19	0.004
	Sliders actions	-0.24	0.01

Table 9. Significant effects found on regressions over subjective metrics applicable only to the controllable interface.

An increased use of the Venn diagram caused a significant increase of user perception of it as a valuable tool as measured by their answer to statements "*I think the Venn diagram visualization helped me to understand why a talk was recommended*" (C_VENN_UNDERSTAND) and "*I think the Venn diagram visualization was useful to identify talks recommended by a specific recommendation method or by a combination of recommendation methods*". More specifically, a unit increase in log-count usage of Venn diagram increased agreement with the first statement by 0.17 and with the second by 0.19. In contrast, the increased use of sliders caused a *decrease* of the same parameters. A unit increase in the log-count usage of sliders produced a decrease on agreement with the first statement by 0.27 units and with the second by 0.24 units. It is interesting, though, that there is no significant effect of either actions on the statement "I understood why the talks were recommended to me." These results suggest that the effect of sliders and the Venn diagram on user perception of the Venn diagram is not necessary stackable. These two tools might compete for user attention potentially causing users to diverge into those who tend to prefer sliders and have a lower attitude to the usefulness of the Venn diagram and those who prefer the Venn diagram and have a higher attitude to it.

6.5. Summary of Quantitative Analysis

The three sets of regression analyses on metrics of different dimensions (usage, objective and subjective) provided answers to the research questions stated in this study. Next we summarize implications on the results to the four research questions.

6.5.1. RQ1. How does controllability affect the user engagement on a recommender system?

The results of the regression analysis on usage metrics (time spent) and on subjective metrics in the post-session survey that assessed different dimensions (perceived usability, endurability, novelty) indicated that the users engaged with the controllable recommender interface, although the effect of the increased engagement is significant only in cases when users have gained some experience with the basic system and the nature of the task during the study. Using SetFusion as second interface, after performing the bookmarking task with the non-controllable interface, significantly increased their engagement time with the systems. As the subjective data showed, the time was spent in an enjoyable

work with the system, not struggling, since subjects' agreement with statements that the controllable user interface was easy to learn, that they would use it again and that they would suggest it to colleagues has also significantly increased in this condition.

6.5.2. RQ2. How does controllability affect the user experience in a recommender system?

As in the discussion of the effects on user engagement, the controllable interface showed a positive effect *without* interaction with other variables in two important metrics: Mean Average Precision (MAP), and understandability of the interface (UNDERSTOOD). The first one, MAP, is an objective metric used frequently in IR, and it showed that the controllable interface does a better job, on average, in bringing relevant items to the top. The other metric is understandability, a subjective metric evaluated in the post-session survey. It showed that the design of the controllable interface actually triggers a better perception about understanding what is being recommended compared to the baseline interface.

Another five subjective metrics showed a positive effect of the treatment (the controllable interface), but the effect reaches significance when the participants used the controllable interface after using the non-controllable one. These metrics are related to a positive perception of the interface ('easy to get familiar with it' and 'overall satisfaction'), a feeling of not missing important talks, and the willingness to use the interface again and to recommend it to other people.

6.5.3. RQ3. Do user characteristics affect the role of controllability on the user engagement with a recommender system?

The analysis shows an important effect of different user characteristics on user engagement with the recommender system. Experience with Conference Navigator, experience with recommender systems, trusting propensity, trust in recommender systems and expertise in the research domain had significant effects on users' engagement.

On the first set of regressions we identified that users were more likely to explore the talks using the traditional actions (checking the abstract rather than exploring the new features) if they had previous experience with the system. On the other side, users with experience in recommender system were less likely to explore the talks in this traditional way (reading the abstract), probably because they were led by the explanatory recommendation features of the controllable interface.

Another interesting finding is the role of trust. As a result of the factor analysis, we considered two types of trust: the general trusting propensity of a user and the more specific trust in recommender systems. However, the second one can be misleading, because it actually measures the trust of users on current and traditional implementations of recommender systems like Amazon.com, or Netflix. The distinction is important because both characteristics had slightly different effects on the user behavior in this study: high trusting propensity makes people have a better appreciation of different forms to deliver recommendations like the Venn diagram, and high trust on recommenders increased the use the sliders, which is a less transparent, but more traditional widget. Although these two visual widgets were designed to be complementary, increased use of the sliders had two effects: (a) participants were less likely to think that the Venn diagram was useful in understanding the fusion of different recommenders, and (b) it decreased agreement that the Venn diagram underpinned users' trust on the recommended list. The second variable playing a role in engagement (the number of actions with the Venn diagram and with the sliders) is *being a native English speaker*. Native English speakers judged talks as relevant or not based more on their content and less on the visual controllable features.

6.5.4. RQ4. Do user characteristics affect the role of controllability on the user experience in a recommender system?

Several user characteristics had important effects on the user experience in the recommender system investigated. A high trusting propensity not only led to the increased perception of the Venn diagram as a useful tool (as discussed above), but also led to an increased perception of the Venn diagram as a tool that increased trust on the recommendations. Gender impacted the perception of understanding why a talk was recommended: males agreed less than females that the Venn diagram helps them understand why a talk was recommended. Native speakers had better perception than non-native English speakers in understanding why the talks where recommended.

Among various individual parameters, past expertise of different kinds appears to be an important factor. The experience with recommender systems had a significant effect on perceiving that Conference Navigator actually needs a recommender system. Participants with more expertise, although less engaged with the system in terms of being immersed in the task, were able to distinguish talks at a fine-grain level, and so they perceived a significantly higher level of diversity in the items recommended (variable DIVERSE in Table 7). A higher perception of item diversity has been associated with a good user experience (Ziegler et al., 2005), since users didn't feel that the recommendations were accurate but *obvious*.

7. Results on Post-Study Survey and Qualitative Analysis

At the end of the study session, each subject answered a post-survey in order to collect the user's perception when comparing the recommender interfaces: the static list and the visual controllable one. A qualitative analysis of the user comments at question six is presented in the next subsection.

Questions one and two asked participants about their preferred interface and which one they would suggest to permanently implement in Conference Navigator. The results of both questions showed a clear preference for the visual controllable interface. For the first question 36 out of 40 participants (90%) preferred the visual controllable recommender and 4 out of 40 (10%) participants liked both interfaces. Regarding which interface they would recommend to implement permanently in Conference Navigator, only one subject would not suggest implementing either the visual controllable or the static list interface, and another subject recommends implementing the static list of recommendations. The other 38 participants would recommend implementing the visual controllable recommender only (33 out of 40, 82.5%) or both interfaces (5 people, 12.5%).

The third question asked about the perceived effort; more particularly, which of the interfaces required more effort to complete the task of finding relevant articles. Although there was no absolutely preferred answer as in the previous two questions, the percentage of users that considered the controllable list of recommendations as requiring more effort to complete the task was only 17.5% compared to 60% of participants who considered the non-controllable as the one requiring more effort. Using a one-sample test of proportions, this 17.5% is significantly different than a null probability of 60%, $\chi^2 = 28.36$, $p < 0.001$.

7.1. Positive comments on the visual interface

7.1.1. Usefulness of Venn diagram intersections

Ten people particularly praised the Venn diagram filtering interaction and its capability to show and filter intersections of algorithms, i.e., papers recommended by more than one method. One user found that most of her selected papers were exactly in the intersections displayed in the Venn diagram, making her task easy:

"I like the Venn diagram especially because most papers I was interested in fell in the same intersections, so it was pretty easy to find and bookmark the relevant papers through it. In the static list I felt almost stressed that I practically had to read all the abstracts to find the papers relevant to me".

Another related comment highlighted the Venn diagram's explainability and another the usefulness of intersections:

"Venn diagram was more helpful as you could actually see the criteria for a given recommendation. Papers in intersection mostly matched choices finally made, they actually matched my interests."

"I like the visual one. It's clear and the Venn diagram figure can help to find relevant information. The intersection of the Venn diagram is very helpful; you won't miss any information through such graph."

7.1.2. Sliders to filter and combine different criteria

Some users preferred the sliders over the Venn diagram. Interestingly, most of these subjects were men, and this observation was supported by the analysis of user characteristics that affected controllability in the previous section. One user commented:

"I had too many things in my head: the sliders, the Venn diagram, the papers I had to find for me and for my colleagues. This made me a bit exhausted and focused on the sliders tool."

The sliders widgets also allowed a fuzzier filtering, which some preferred:

"I like the visual controller since I can determine the combination of criterion between my preferences, conferences attendees and authors' reputation. I preferred the sliders over the Venn diagram"

Another important point that was missed in the surveys was the familiarity of the users with some visual interaction methods.

"I prefer the sliders because I have used a system before to control search results with a similar widget, so I was more familiar to me."

7.1.3. Transparency and explainability of the Visual controllable interface

The fact that the controllable interface provided explanations and clear criteria of why the papers were recommended made people prefer it over the non-controllable interface. One user commented:

" (the visual interface) … made me feel that there was a reason for the articles being presented that I could control, rather than the reason controlled by some unknown algorithms…"

Another user commented on the same characteristics, and pointed out that this increased his confidence in the results:

"I prefer the visual controllable recommender (VCR) to the static list of recommendations. VCR provided intuition to understand why the items were recommended. This increased my confidence on the results suggested."

Another user was explicit in missing the transparency in the non-controllable interface:

"It was much less evident to me why I was presented with the recommendations in the static list. I appreciated the transparency in the controllable interface."

In addition to transparency and explainability, several users found the controllable interface more appealing and engaging. Two people commented that the interface was easy to use and the controllability capabilities engaged them:

"Visual interface: easy to learn how to use, easy to sort based on my preferences (sliders) with a clear interface."

"the visual controllable interface makes me more confident that I am not missing interesting articles and made looking for them not so boring"

7.2. Critical comments on the visual interface

7.2.1. Venn diagram is redundant

One unexpected finding from the analysis of user characteristics that could affect user experience was that being male increased the odds of interacting more with the sliders. Indeed, most people that answered that they preferred the sliders over the Venn diagram at the end of the study were men. Some of the reasons they gave:

"Don't like the Venn diagram, is redundant. Sliders and colors by item were enough to tell the relevancy of an item."

"By the time of using the controllable interface I was a bit tired and I focused on the task of finding relevant papers rather than exploring all the capabilities of the interface. The sliders were very efficient in helping me to filter papers by different criteria and find the relevant ones."

One characteristic of the sliders widget is that it can reproduce one of the filtering capabilities of the Venn diagram. Setting one of the bars to weight 1 and the other two sliders to 0 was equivalent to clicking on some of the areas of the Venn diagram to filter, while also showing other papers at the bottom of the list.

7.2.2. Short list of items makes visual widgets unnecessary

"I thought the controllable one adds unnecessary complication if the list is not very long"

Three users commented that they would have found the visual widgets more useful if the list of recommended items was longer. Since the recommendation lists were at most 60 articles, these subjects found that it was easier simply scanning through all the papers than learning to use visual features. This is a very important observation, since recommender systems are supposed to be an important aid in helping users to filter large amounts of information, and in this study subjects had to find only 15 relevant papers out of 60.

7.2.3. Affordance

One user commented that there was a disconnection in the sliders that made him feel that there is a potential improvement in the interface:

"*at one point I filtered the papers using the Venn diagram, then I reset the weights with the sliders and clicked in the update button. I was surprised that my filter on the Venn diagram was focused and even confused me…*"

This is actually a design feature of the controllable interface and it should be made clearer to users how it works to avoid confusion or lose of trust in the interface.

One user commented that he would like to be able to tell the system that there are one or more items that should be removed. This same comment was made after the CSCW study (Parra & Brusilovsky, 2013), and it should be considered for the next version of this system.

8. Discussion and Conclusions

This paper investigated the prospects of visual interface for user-controllable personalization in the context of hybrid recommender systems. We introduced a novel recommendation interface, SetFusion, that uses Venn diagram visualization, slider-based fusion controllability, and some other features to support controllability and transparency of a hybrid recommendation process. One of the reasons to pick this kind of visualization, which has not been used so far to visualize recommendations, is that it provides "different depths of field" defined by Lurie and Mason (2007) as "the extent to which a visualization provides contextual overview versus detail information or enable decision makers to keep both levels in focus at the same time". Using the Venn diagram to explain intersections among recommendation approaches, i.e. what items they have in common and which are recommended by only one method, provides different depths of field, a positive characteristic in decision making that "… allows the user to focus on a subset of alternatives but remain cognizant of others" (Lurie & Mason, 2007). The novel Venn diagram feature has been combined with an extended slider-based controls that have been aleady suggested (Bostandjiev et al., 2012) but remain underexplored.

In order to assess the impact of our proposed interface, we installed SetFusion as a component of the CN3 conference recommendation system and conducted a user study where SetFusion was compared with a baseline "ranked list" recommender interface in the task of finding interesting conference talks. The data obtained by the study has been analyzed in several ways including an extensive regression analysis where we used three dimensions of metrics (behavioral, objective and subjective) and controlled for user characteristics to obtain a better understanding of factors influencing the acceptance of our recommender. The results of this work present an interesting picture of the overall influence of the interface as well as an interaction effect with the order in which the interfaces were presented, and several user characteristics.

8.1. Overall Effect

Starting with the overall effect, one thing that we observed is an extensive use of various features provided by the SetFusion interface – 32 out of 40 users used the Venn diagram and 33 used sliders. This data provides evidence that the system was considered worth trying by the dominated majority of the users. The average amount of use was quite considerable and the effect was apparently positive. The

significant increase of MAP in the SetFusion condition indicates that ranked lists obtained after slider-based re-ranking or filtering presented better talks to the users. The bookmarking data shows that the use of both tools was not in vain – the subjects made the vast majority of bookmarks immediately after using one of the provided tools. SetFusion also significantly increased user understanding of why talks were recommended. We believe that it could be mainly attributed to the presence of the Venn diagram, which was engineered to deliver this information. An additional effect of the Venn diagram specifically appreciated by the users was its ability to attract user attention to interesting talks in the overlapping areas containing talks recommended by more than one method. The analysis of user ratings for different areas of the Venn diagram demonstrated that user appreciation had a solid reason – talks in the overlapped areas tend to be rated higher by the users demonstrating their increased relevance. All of these positive features of SetFusion resulted in an overall positive feedback. All 40 participants selected SetFusion as their preferred interface and 38 participants selected is the one to be implemented permanently as a part of the Conference Navigator.

At the same time, the study data indicated that the SetFusion interface was not working equally well for all participants. The usage table showed that some participants had little or no use of its advanced features. Several participants, while positive about SetFusion, were not ready to give away the simpler baseline version – 4 of them selected baseline as the preferred interface along with SetFusion, five wanted to have the baseline interface implemented in Conference Navigator along with SetFusion and one suggested to implement the baseline only. The question about the perceived effort showed the same divided picture: while 60% of participants considered the non-controllable baseline as the interface requiring more efforts, still 17.5% believed that it was SetFusion that required more efforts. The analysis of study data hinted that this split of opinions was a result of relative simplicity of the task contrasted with a relative complexity of the interface. On one hand, several users acknowledged that given a rather small number of talks in the ranked list, there was little need in any advanced tools. On the other hand, a combination of a new system (CN) and new context (talk recommendation) provided a considerable cognitive overhead discouraging some users from using two new tools (sliders and Venn diagram) on the top of that.

In this context, the use of SetFusion and the preference for it were defined by a careful balance of user needs, skills, and motivation. For example, native speakers who could more easily scan a list of 30 talks without further re-ranking or filtering performed significantly fewer actions with both the Venn diagram and sliders. With or without SetFusion they also better understood why the talks were recommended. In contrast, non-native speakers needed all support they can get in finding good talks and were more willing to engage SetFusion. A similar factor working in the opposite direction was higher trust in recommender systems. The increased trust apparently provided additional motivation to invest time in learning SetFusion increasing the number of actions on the sliders by 44%.

Most visible among factors that were able to affect the balance in favor of SetFusion or against it were various experience factors. We can argue that having previous experience in any components of the overall tasks could decrease the total cognitive load leaving more space for exploring SetFusion features. For example, being already familiar with the conference system significantly increased the number of SetFusion actions. The easiest way to gain more experience in systems components was the use of the baseline interface before SetFusion. A simpler baseline interface worked as training wheels for SetFusion allowing the users to get familiar with the task, the overall systems and some interface components. As a result, we observed several effects of interface order. The application of SetFusion as the second interface led to a 82% increase in using sliders. It also caused a significant improvement of

user perception about SetFusion on five out of 10 metrics. The data also indicates that sufficient experience that can affect user perception of the interface could be gained even within the session. For example, an increased use of the Venn diagram caused a significant increase of user perception of it as a valuable tool.

An additional support for the "cognitive load" factor is provided by unexpected competition between the sliders and the Venn diagram that we observed. Most clearly this competition was stressed in user feedback – several users indicated that in some aspects these two tools were redundant – some effect (like filtering out one source) could be achieved with either of the tools. The regression results supported this observation showing that the increased use of the sliders decreased user opinion about usefulness of the Venn diagram. Apparently, it was quite a challenge to master both tools, thus some users focused on sliders at the expense of the Venn diagram and vice versa. We do not think, however, that the use of both tools in SetFusion is a negative factor. In contrast, it provided more flexibility allowing users to focus on a tool that is more familiar or attractive for them.

The effect of experience on the value and perception of the innovative interfaces is not new – it is rather typical for a range of interfaces from advanced search to advanced personalization. However, we believe that the increase of task complexity could make the value of controllability and transparency, the cornerstones of the SetFusion interface more pronounced for all users regardless of their past experience. We hope to assess this belief in the future studies. We also plan a more extensive exploration of various user features that affect user experience with SetFusion and its components. In particular, we are interested to explore which user features make the sliders or the Venn diagram more attractive – the findings in this area could lead to interface-level adaptation. We also plan to explore further the impact of some less evident factors such as gender on the usage of SetFusion and its perception. We also hope that similar research of other teams on the transparency and controllability of recommender interfaces will help us better understand the impact of these factors eventually leading to better recommender systems.

9. References

Albers, M. J. 1997. *Cognitive strain as a factor in effective document design*. Paper presented at the Proceedings of the 15th annual international conference on Computer documentation, Salt Lake City, Utah, USA.

Attfield, S., Kazai, G., Lalmas, M., & Piwowarski, B. 2011. Towards a science of user engagement *WSDM Workshop on User Modeling for Web Applications*.

Bennett, J., Lanning, S., & Netflix, N. 2007. *The Netflix Prize.* Paper presented at the In KDD Cup and Workshop in conjunction with KDD.

Bostandjiev, S., O'Donovan, J., & Höllerer, T. 2012. *TasteWeights: a visual interactive hybrid recommender system.* Paper presented at the Proceedings of the sixth ACM conference on Recommender systems, New York, NY, USA.

Bostandjiev, S., O'Donovan, J., & Höllerer, T. 2013. *LinkedVis: exploring social and semantic career recommendations*. Paper presented at the Proceedings of the 2013 international conference on Intelligent user interfaces, Santa Monica, California, USA.

Brusilovsky, P., Parra, D., Sahebi, S., & Wongchokprasitti, C. 2010. *Collaborative information finding in smaller communities: The case of research talks.* Paper presented at the CollaborateCom.

Burke, R. 2002. Hybrid Recommender Systems: Survey and Experiments. *User Modeling and User-Adapted Interaction, 12*(4), 331-370.

Cramer, H., Evers, V., Ramlal, S., van Someren, M., Rutledge, L., Stash, N., Wielinga, B. 2008. The effects of transparency on trust in and acceptance of a content-based art recommender. *User Modeling and User-Adapted Interaction, 18*, 455-496.

Goldberg, D., Nichols, D., Oki, B. M., & Terry, D. 1992. Using collaborative filtering to weave an information tapestry. *Commun. ACM, 35*, 61-70.

Gretarsson, B., O'Donovan, J., Bostandjiev, S., Hall, C., & Höllerer, T. 2010. *Smallworlds: visualizing social recommendations*. Paper presented at the Proceedings of the 12th Eurographics / IEEE - VGTC conference on Visualization, Bordeaux, France.

Herlocker, J. L., Konstan, J. A., & Riedl, J. 2000. *Explaining collaborative filtering recommendations.* Paper presented at the Proceedings of the 2000 ACM conference on Computer supported cooperative work, New York, NY, USA.

Hijikata, Y., Kai, Y., & Nishida, S. 2012. *The relation between user intervention and user satisfaction for information recommendation.* Paper presented at the Proceedings of the 27th Annual ACM Symposium on Applied Computing, New York, NY, USA.

Hu, Y., Koren, Y., & Volinsky, C. 2008. *Collaborative Filtering for Implicit Feedback Datasets.* Paper presented at the Proceedings of the 2008 Eighth IEEE International Conference on Data Mining, Washington, DC, USA.

Jameson, A., & Schwarzkopf, E. 2006. *Pros and cons of controllability: An empirical study.* Paper presented at the Adaptive Hypermedia and Adaptive Web-based Systems.

Järvelin, K., Price, S., Delcambre, L. L., & Nielsen, M. 2008. Discounted Cumulated Gain Based Evaluation of Multiple-Query IR Sessions. In C. Macdonald, I. Ounis, V. Plachouras, I. Ruthven & R. White (Eds.), *Advances in Information Retrieval* (Vol. 4956, pp. 4-15): Springer Berlin Heidelberg.

Järvelin, K., Price, S. L., Delcambre, L. M. L., & Nielsen, M. L. 2008. Discounted Cumulated Gain Based Evaluation of Multiple-Query IR Sessions. In C. Macdonald, I. Ounis, V. Plachouras, I. Ruthven & R. White (Eds.), *Advances in Information Retrieval* (Vol. 4956, pp. 4-15): Springer Berlin Heidelberg.

Knijnenburg, B. P., Bostandjiev, S., O'Donovan, J., & Kobsa, A. 2012. *Inspectability and control in social recommenders.* Paper presented at the Proceedings of the sixth ACM conference on Recommender systems, New York, NY, USA.

Knijnenburg, B. P., Reijmer, N. J. M., & Willemsen, M. C. 2011. *Each to his own: how different users call for different interaction methods in recommender systems.* Paper presented at the Proceedings of the fifth ACM conference on Recommender systems, New York, NY, USA.

Knijnenburg, B. P., Willemsen, M. C., Gantner, Z., Soncu, H., & Newell, C. 2012. Explaining the user experience of recommender systems. *User Modeling and User-Adapted Interaction, 22*(4-5), 441-504.

Lurie, N. H., & Mason, C. H. 2007. Visual Representation: Implications for Decision Making. *Journal of Marketing, 71*, 160-177.

Manning, C. D., Raghavan, P., & Schtze, H. 2008. *Introduction to Information Retrieval*. New York, NY, USA: Cambridge University Press.

Marujo, L., Bugalho, M., Neto, J. P. d. S., Gershman, A., & Carbonell, J. 2013. Hourly Traffic Prediction of News Stories. *arXiv preprint arXiv:1306.4608*.

McNee, S. M., Riedl, J., & Konstan, J. A. 2006a. *Being accurate is not enough: how accuracy metrics have hurt recommender systems.* Paper presented at the CHI '06 extended abstracts on Human factors in computing systems, New York, NY, USA.

McNee, S. M., Riedl, J., & Konstan, J. A. 2006b. *Making recommendations better: an analytic model for human-recommender interaction.* Paper presented at the CHI '06 Extended Abstracts on Human Factors in Computing Systems, New York, NY, USA.

O'Brien, H. L., & Toms, E. G. 2010. The Development and Evaluation of a Survey to Measure User Engagement. *J. Am. Soc. Inf. Sci. Technol., 61*(1), 50-69.

O'Donovan, J., Smyth, B., Gretarsson, B., Bostandjiev, S., & Höllerer, T. 2008. *PeerChooser: visual interactive recommendation.* Paper presented at the Proceedings of the twenty-sixth annual SIGCHI conference on Human factors in computing systems, New York, NY, USA.

Parra, D., & Brusilovsky, P. 2013. *A field study of a visual controllable talk recommender*. Paper presented at the Proceedings of the 2013 Chilean Conference on Human - Computer Interaction, Temuco, Chile.

Parra, D., Brusilovsky, P., & Trattner, C. 2014. *See what you want to see: visual user-driven approach for hybrid recommendation*. Paper presented at the Proceedings of the 19th international conference on Intelligent User Interfaces, Haifa, Israel.

Parra, D., Jeng, W., Brusilovsky, P., López, C., & Sahebi, S. 2012. *Conference Navigator 3: An Online Social Conference Support System.* Paper presented at the Workshop and Poster Proceedings of the 20th Conference on User Modeling, Adaptation, and Personalization Montreal, Canada, July 16-20, 2012.

Parra, D., & Sahebi, S. 2013. Recommender Systems: Sources of Knowledge and Evaluation Metrics. In J. D. V. a. squez & et al. (Eds.), *Advanced Techniques in Web Intelligence-2: Web User Browsing Behaviour and Preference Analysis* (pp. 149-175). Berlin Heidelberg: Springer-Verlag.

Pu, P., Chen, L., & Hu, R. 2011. *A user-centric evaluation framework for recommender systems.* Paper presented at the Proceedings of the fifth ACM conference on Recommender systems, New York, NY, USA.

Resnick, P., Iacovou, N., Suchak, M., Bergstrom, P., & Riedl, J. 1994. *GroupLens: an open architecture for collaborative filtering of netnews.* Paper presented at the Proceedings of the 1994 ACM conference on Computer supported cooperative work, New York, NY, USA.

Shardanand, U., & Maes, P. 1995. *Social information filtering: algorithms for automating word of mouth.* Paper presented at the Proceedings of the SIGCHI conference on Human factors in computing systems, New York, NY, USA.

Sherman, E. H., & Shortliffe, E. H. 1993. A User-Adaptable Interface to Predict Users' Needs. In M. Schneider-Hufschmidt, T. Kuhme & U. Malinowski (Eds.), *Adaptive User Interfaces: Principles and Practice* (pp. 285-316). Amsterdam: North-Holland.

Tintarev, N., & Masthoff, J. 2007. *Effective explanations of recommendations: user-centered design.* Paper presented at the Proceedings of the 2007 ACM conference on Recommender systems, New York, NY, USA.

Tintarev, N., & Masthoff, J. 2011. Designing and Evaluating Explanations for Recommender Systems. In F. Ricci, L. Rokach, B. Shapira & P. B. Kantor (Eds.), *Recommender Systems Handbook* (pp. 479-510): Springer US.

Verbert, K., Parra, D., Brusilovsky, P., & Duval, E. 2013. *Visualizing recommendations to support exploration, transparency and controllability*. Paper presented at the Proceedings of the 2013 international conference on Intelligent user interfaces, Santa Monica, California, USA.

Ziegler, C.-N., McNee, S. M., Konstan, J. A., & Lausen, G. 2005. *Improving recommendation lists through topic diversification.* Paper presented at the Proceedings of the 14th international conference on World Wide Web, New York, NY, USA.

10. Vitae

Denis Parra works in recommender systems, intelligent user interfaces, and social media analysis. He has published papers in several conferences in the area of recommender systems and personalization such as RecSys, IUI, and UMAP. In this last conference, he earned a best student paper award in 2011 with his research on implicit feedback recommendation. Denis currently works as Assistant Professor in the School of Computer Science at Pontificia Universidad Catolica de Chile. He holds a B.Eng. degree from Universidad Austral de Chile and a Ph.D. degree in Information Science from the University of Pittsburgh.

Peter Brusilovsky has been working in the field of adaptive educational systems, user modeling, and intelligent user interfaces for more than 20 years. He published numerous papers and edited several books on adaptive hypermedia and the adaptive Web. Peter is currently Professor of Information Science and Intelligent Systems at the University of Pittsburgh, where he directs Personalized Adaptive Web Systems (PAWS) lab. Peter is the Associate Editor-in-Chief of IEEE Transactions on Learning Technologies and a board member of several journals including User Modeling and User Adapted Interaction, ACM Transactions on the Web, and Web Intelligence and Agent Systems.

11. Appendix

11.1. Pre-Study Survey

1. Current degree/program or position: _____

2. Gender: ___ Male ___Female

3. Are you a native English speaker: ___ Yes ___No

4. Age: _____

The Following 9 questions are related to your experience as researcher, your familiarity with the iConference, and your familiarity with Conference Navigator

5. If you are pursuing a PhD degree, which stages have you completed in your program of study?

Preliminary Exam / Comprehensive Exam / Proposal Defense / Dissertation Defense

6. How many workshop papers or posters have you published in your area of research?

None / 1-2 / 3-4 / 5 or more

7. How many conference or journal papers have you published in your area of research?

None / 1-2 / 3-4 / 5 or more

8. Have you served as a reviewer for workshops, conferences or journals in your area of research?

Yes / No

---

9. How many iConferences have you attended?

None / one / 2-4 / 5 or more

10. How many papers have you published in the iConference?

None / one / 2-4 / 5 or more

11. I feel engaged with the iSchools community

Strongly disagree / Disagree / Neutral / Agree / Strongly Agree

---

12. I have used Conference Navigator in the past

Yes / No

13. I am familiar with the features of Conference Navigator

Strongly disagree / Disagree / Neutral / Agree / Strongly Agree

In the next questions, answer how much do you agree with the following statements

14 In general, people really do care about the well-being of others.

Strongly disagree / Disagree / Neutral / Agree / Strongly Agree

15. The typical person is sincerely concerned about the problems of others.

Strongly disagree / Disagree / Neutral / Agree / Strongly Agree

16. Most of the time, people care enough to try to be helpful, rather than just looking out for themselves.

Strongly disagree / Disagree / Neutral / Agree / Strongly Agree

17. I am familiar with online recommender systems.

Strongly disagree / Disagree / Neutral / Agree / Strongly Agree

18. I have occasionally followed the advice of a recommender system (such as a recommended book in Amazon.com or a recommended video in YouTube)

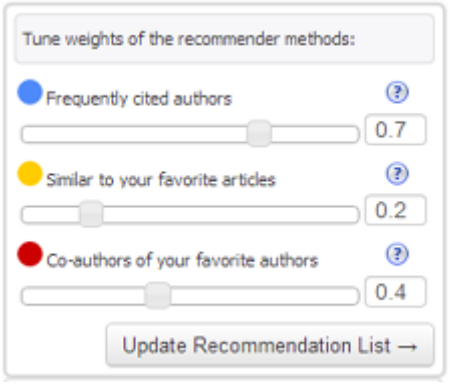
Strongly disagree / Disagree / Neutral / Agree / Strongly Agree

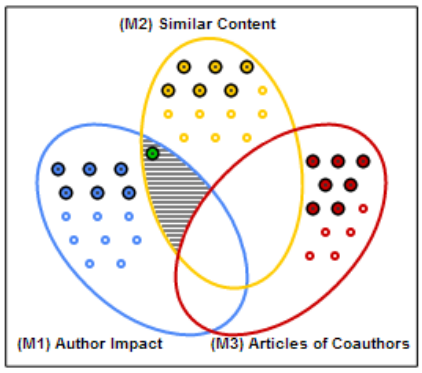
19. I know of one or more methods used to produce recommendations in a system

Strongly disagree / Disagree / Neutral / Agree / Strongly Agree

11.2. Post-session survey (controllable condition)

Talks were recommended based on three different recommendation methods. The current interface was designed to allow users to manipulate the importance of each of the recommendation methods by using sliders, and to examine the items recommended by each method using a Venn diagram.

<table>
<tr>
<td colspan="2"> 

Sliders</td>
<td colspan="2"> 

Venn Diagram</td>
</tr>
<tr><td></td><td colspan="3"><< To what extent do you agree with the following statements? >> (items with * apply only to users in the controllable interface condition)</td></tr>
<tr><td>1</td><td>UNDERSTOOD</td><td colspan="2">I understood why the talks were recommended to me.</td></tr>
<tr><td>2</td><td>RELEVANT</td><td colspan="2">The items recommended matched my interests.</td></tr>
<tr><td>3</td><td>DIVERSE</td><td colspan="2">The items recommended were diverse.</td></tr>
<tr><td>4</td><td>INTERFACE_EASY</td><td colspan="2">I became familiar with the recommender interface very quickly</td></tr>
<tr><td>5</td><td>LOST_TRACK_TIME</td><td colspan="2">I lost track of time while I was using the recommender interface</td></tr>
<tr><td>6</td><td>OVERALL_SATISFIED</td><td colspan="2">Overall, I am satisfied with the recommender interface</td></tr>
<tr><td>7</td><td>CONFIDENT_MISS</td><td colspan="2">The recommender made me more confident that I didn’t miss relevant talks</td></tr>
<tr><td>8</td><td>USE_AGAIN</td><td colspan="2">I would use this recommender system again for another conference in the future</td></tr>
</table>

9	SUGGEST_COLLEAGUES	I would suggest my colleagues to use this recommender system when they attend a conference in the future
10	RECSYS_NO_NEED	I do not think that a social conference support system - like Conference Navigator- needs Talk Recommendation functionality
	Statements referring only to the controllable interface	
*1	C_FEEL_CONTROL	I felt in control of combining different recommendation methods by using the sliders.
*2	C_ABIL_CONT_SATISF	The ability to control the recommendation methods increased my satisfaction with the list of recommended talks.
*3	C_ABIL_CONT_TRUST	The ability to control the recommendation methods increases my trust in the list of recommended talks.
*4	C_INTEREST_EXAMINE	When looking at the list of recommended talks I am interested to examine which recommendation method has been used.
*5	C_VENN_UNDERSTAND	I think the Venn diagram visualization helped me to understand why a talk was recommended.
*6	C_VENN_USE	I think the Venn diagram visualization was useful to identify talks recommended by a specific recommendation method or by a combination of recommendation methods.
*7	C_VENN_TRUST	The ability to use the Venn diagram to examine the talks recommended increases my trust in the list of recommended talks.

11.3. Post-session survey (Non-controllable condition)

<< To what extent do you agree with the following statements? >>		
1	UNDERSTOOD	I understood why the talks were recommended to me.
2	RELEVANT	The items recommended matched my interests.
3	DIVERSE	The items recommended were diverse.
4	INTERFACE_EASY	I became familiar with the recommender interface very quickly
5	LOST_TRACK_TIME	I lost track of time while I was using the recommender interface
6	OVERALL_SATISFIED	Overall, I am satisfied with the recommender interface
7	CONFIDENT_MISS	The recommender made me more confident that I didn't miss relevant talks
8	USE_AGAIN	I would use this recommender system again for another conference in the future

9	SUGGEST_COLLEAGUES	I would suggest my colleagues to use this recommender system when they attend a conference in the future
10	RECSYS_NO_NEED	I do not think that a social conference support system - like Conference Navigator- needs Talk Recommendation functionality
11	Comments and general feedback from the subject.	

11.4. Post- Study survey

This survey was used only in the study 2 (iConference study)

1. Which one of the interfaces did you like/prefer most?

a) The static list of recommendations
b) The visual controllable recommender
d) I liked both of them
c) I didn't like any of them

2. Which of the interfaces would you suggest to implement permanently in Conference Navigator?

a) The static list of recommendations
b) The visual controllable recommender
c) I wouldn't suggest to implement any of them
d) I would suggest to implement both of them

3. Which of the interfaces did you feel that required more effort in order to find relevant articles?

a) The static list of recommendations
b) The visual controllable recommender
c) Both required more or less the same level of effort
d) I cannot tell which one required more effort

4. Overall how would you rate the static list recommendations interface?

1 (I don't like it at all) 2 3 (I don't know) 4 5 (I really like it)

5. Overall how would you rate the visual controllable recommendation interface?

1 (I don't like it at all) 2 3 (I don't know) 4 5 (I really like it)

6. In case that you preferred one interface over the other, could you elaborate about your preference? [This answer can be talk-aloud and would be recorded]