USER CONTROLLABILITY IN A HYBRID RECOMMENDER SYSTEM

by

Denis Alejandro Parra Santander

B.S Engineering Sciences, Universidad Austral de Chile, 2002

Submitted to the Graduate Faculty of

The School of information Sciences in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2013

UNIVERSITY OF PITTSBURGH

School of Information Sciences

This dissertation was presented

by

Denis Alejandro Parra Santander

It was defended on

July 22nd, 2013

and approved by

Peter Brusilovsky, PhD, Professor

Rosta Farzan, PhD, Assistant Professor

Stephen Hirtle, PhD, Professor

Paul Resnick, PhD, Professor

Michael Spring, PhD, Associate Professor

Dissertation Advisor: Peter Brusilovsky, PhD, Professor

USER CONTROLLABILITY IN A HYBRID RECOMMENDER SYSTEM

Denis Parra

University of Pittsburgh, 2013

Since the introduction of Tapestry in 1990, research on recommender systems has traditionally focused on the development of algorithms whose goal is to increase the accuracy of predicting users' taste based on historical data. In the last decade, this research has diversified, with *human factors* being one area that has received increased attention. Users' characteristics, such as trusting propensity and interest in a domain, or systems' characteristics, such as explainability and transparency, have been shown to have an effect on improving the user experience with a recommender. This dissertation investigates on the role of controllability and user characteristics upon the engagement and experience of users of a hybrid recommender system. A hybrid recommender is a system that integrates the results of different algorithms to produce a single set of recommendations. This research examines whether allowing the user to control the process of fusing or integrating different algorithms (i.e., different sources of relevance) results in increased engagement and a better user experience. The essential contribution of this dissertation is an extensive study of controllability in a hybrid fusion scenario. In particular, the introduction of an interactive Venn diagram visualization, combined with sliders explored in a previous work, can provide an efficient visual paradigm for information filtering with a hybrid recommender that fuses different prospects of relevance with overlapping recommended items. This dissertation also provides a three-fold evaluation of the user experience: objective metrics, subjective user perception, and behavioral measures.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

PREFACE

Thanks to life for giving me a second chance. Thanks to Dr. Alex Hills and to my advisor Prof. Peter Brusilovsky for believing in my capacities without even knowing me.

Thanks to my parents Zuny for teaching me perseverance, and Sergio for teaching me to be patient. Thanks to my sisters Marcia and Soledad. Thanks to my daughter Ariadna, my main inspiration, and thanks Monica for being a great mother and a hard-working woman.

Thanks to Julio Daniel Guerra for being such a good friend for almost 20 years. Thanks to Fernando Cerda for those endless talks to fix, from Pittsburgh, the problems of our beautiful Chile. Thanks to Jonathan Jeong, my 5-year roommate that has taught me that human beings can be honestly generous and disinterested. Thanks to all my Pittsburgh family, the PAWS lab: Rosta, Mike, Sergey, Sharon, Jenn, … sorry no more space for listing all of you guys. Thanks to Xavier Amatrian. You are one of the greatest persons I have ever worked with. My main objective in life is being as efficient as you, keeping always the strength to pursue my dreams.

Thanks to Shuguang Han for providing the ACM dataset, it saved me countless hours of work. My appreciation also goes to Chris Williams from University of Hawaii for providing the foundational JavaScript code to implement the interactive Venn diagram recommender. Finally, special thanks to Youyang Hou and SangSeok You for their support and feedback during the CSCW study in San Antonio.

1.0 INTRODUCTION

This dissertation investigates HCI (Human-Computer Interaction) aspects of recommender systems. As their name suggests, these are systems that help a user or users to choose items from a large item or information space (McNee, Riedl, & Konstan, 2006a) by proactively suggesting 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 the introduction of attractive 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 the research and development of recommender systems. Historically, the focus on recommender systems' research has been on improving the algorithms' predictive accuracy (D. Parra & Sahebi, 2013), but as McNee et al. highlight in the paper "*Being accurate is not enough: how accuracy metrics have hurt recommender systems*" (McNee, Riedl, & Konstan, 2006b), accuracy does not always correlate with a good user experience, making the study of recommender interfaces one of the areas in need of improvement. With respect to the study of richer 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. User studies on both systems showed increased user satisfaction under the visual interactive interface compared to a more static condition. More recently, TasteWeights (Bostandjiev, O'Donovan, & Höllerer, 2012) implements a visual interactive interface for a hybrid music recommender system. A user study shows that users perceived higher system accuracy under the condition with all the visual and interactive features compared to conditions with only part of the proposed interface.

A second challenge that has been partially addressed, but still has room for further exploration, is how to combine different recommender methods or different contexts of relevance to improve the recommendations generated by a single method. Currently, there are many successful recommender methods, such as collaborative filtering (Goldberg et al., 1992), or content-based (Pazzani & Billsus, 2007), but all of them have their own advantages and disadvantages. Hybrid recommenders (Burke, 2002) are systems that combine different methods of recommendation to overcome their disadvantages; for instance, a recommender using collaborative filtering that is suffering from a sparse dataset can be combined with a content-based recommender or a non-personalized popularity-based recommender in the early stages of such a system. The usual way to combine recommendations of different methods in a hybrid recommender is by automatic off-line optimization; however, some studies show that allowing users to manually control these combinations can be beneficial for their user experience (Bostandjiev et al., 2012; Verbert, Parra, Brusilovsky, & Duval, 2013). In the first study, users are allowed to control the importance of each method with a visual slider interface, whereas the second one shows clickable intersections of different recommender methods by connected clusters of items. Since there are different ways to allow users to intervene and control a recommender system, this dissertation investigates the role of controllability on improving the user experience in a hybrid recommender by combining the successful visual approach of sliders (Bostandjiev et al., 2012; Bostandjiev, O'Donovan, & Höllerer, 2013) with a novel interactive Venn diagram visualization.

In addition to controllability, previous studies have shown the important influence of user characteristics on the user experience with a recommender system. Hijikata et al. show that the influence of controllability on improving the user experience with a recommender system is consistently positive only when the user is highly interested in the domain (Hijikata, Kai, & Nishida, 2012). Another study on an energy-saving recommender system shows that trusting propensity increased the likelihood of users accepting recommendations, and that users with higher domain expertise express higher satisfaction when the system has more control than users with low domain expertise (Knijnenburg, Reijmer, & Willemsen, 2011).

1.1 FOCUS OF THE STUDY

This dissertation investigates how controllability affects the user experience in a recommender system. In order to contribute to previous research on this area, I introduce a novel interface that allows users to control the fusion of different recommendation methods by using visual interactive widgets. This area of research – studying the effect of controllability supported by visual artifacts on the user experience- has been recently pioneered by (O'Donovan et al., 2008) with PeerChooser, (Gretarsson et al., 2010) with SmallWorlds and more recently with TasteWeights (Bostandjiev et al., 2012). The first two studies (on PeerChooser and SmallWorlds) show increased user satisfaction with the visual interactive interface, and the third study (on TasteWeights) showed an increase in users' perception of the system's accuracy on the recommendations. This dissertation extends past work on user controllability in two important areas: a) It introduces a new way to inspect and control a fusion of recommendations (hybrid recommender) through a Venn diagram visualization, inspired by our recent results in (Verbert et al., 2013), and b) It explains the effect of controllability and user characteristics on the user experience by using objective, subjective and behavioral measures. The 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 the available widgets.

Studying recommendations at conferences is important because they depart from more traditional recommendation domains like movies or books. Recommending talks at conferences suffers particularly from *cold-start* and *new-item* problems (users that provide little feedback, items with little information on user preference before the event takes place) since there is only short time from when items are made available (conference program is made available online) to when they are needed (the conference itself), and the items are only valuable for a short period of time: after the articles are presented, they have little value for users in the context of the conference.

Implementing and studying a visual interactive interface with a recommender system to examine the role of controllability on user engagement and on the overall user experience has interesting implications. One should consider the trade-off between being able to manipulate the recommendations and the cognitive burden of having too many potential user-interaction options to make a decision on which items are relevant: a) on the one side, there is an expected increase in user satisfaction with the system since control would provide a better sense of transparency and explainability of the items being recommended, b) this also involves some cognitive burden to the user: instead of having ready-to-choose items, the user must manipulate the interface to make a decision. In some cases, this can be counterproductive for the users' satisfaction.

In order to address these challenges, I have built a novel controllable talk recommendation interface that extends the existent system Conference Navigator, an online web platform that support attendees and organizers of academic conferences. Using this system, I have conducted two user studies to investigate the effect of user controllability on the user experience of a recommender system. The first study compares two interfaces –a traditional static list of recommendations (baseline) against a visual interface with controllable features– in a real conference setting (CSCW 2013); the second study also compares a slightly enhanced version of the visual interface with the baseline in a controlled laboratory setting to study the impact of user characteristics. In addition, I present the results of a field study that was performed to observe how the enhanced visual interface designed for the second study is used in a realistic conference setting.

1.2 OVERVIEW OF THE RESEARCH QUESTIONS

After introducing the importance of human factors in recommender systems, and the focus of this research, in this section I summarize the research questions addressed in this dissertation.

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

This question is 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, O'Donovan, & Kobsa, 2012) that support the effect of controllability on user engagement, but mainly by considering subjective measures (surveys) and one or very few objective metrics (average rating). This dissertation intends to bridge this gap by studying controllability and evaluating user engagement more comprehensively than previous research.

1.2.2 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, Chen, & Hu, 2011) and (Knijnenburg, Willemsen, Gantner, Soncu, & Newell, 2012). Though this dissertation does not present an evaluation using strictly the aforementioned frameworks, it considers and adapts all their dimensions to assess the user experience with subjective metrics. Moreover, this dissertation also considers objective metrics and behavioral measures to complement the evaluation, thus implementing a comprehensive evaluation.

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

Previous research has shown that user characteristics affect engagement with online systems. For instance, O’Brien and Toms develop a construct for evaluating user engagement, novelty, based on the state-trait curiosity model that considers external and internal stimuli, as well as individual differences (O'Brien & Toms, 2010). On the other side, Attfield et al. (2011) survey 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 (Attfield et al., 2011). Investigating whether these results apply in the same way to recommender systems by providing user controllability motivates this research question.

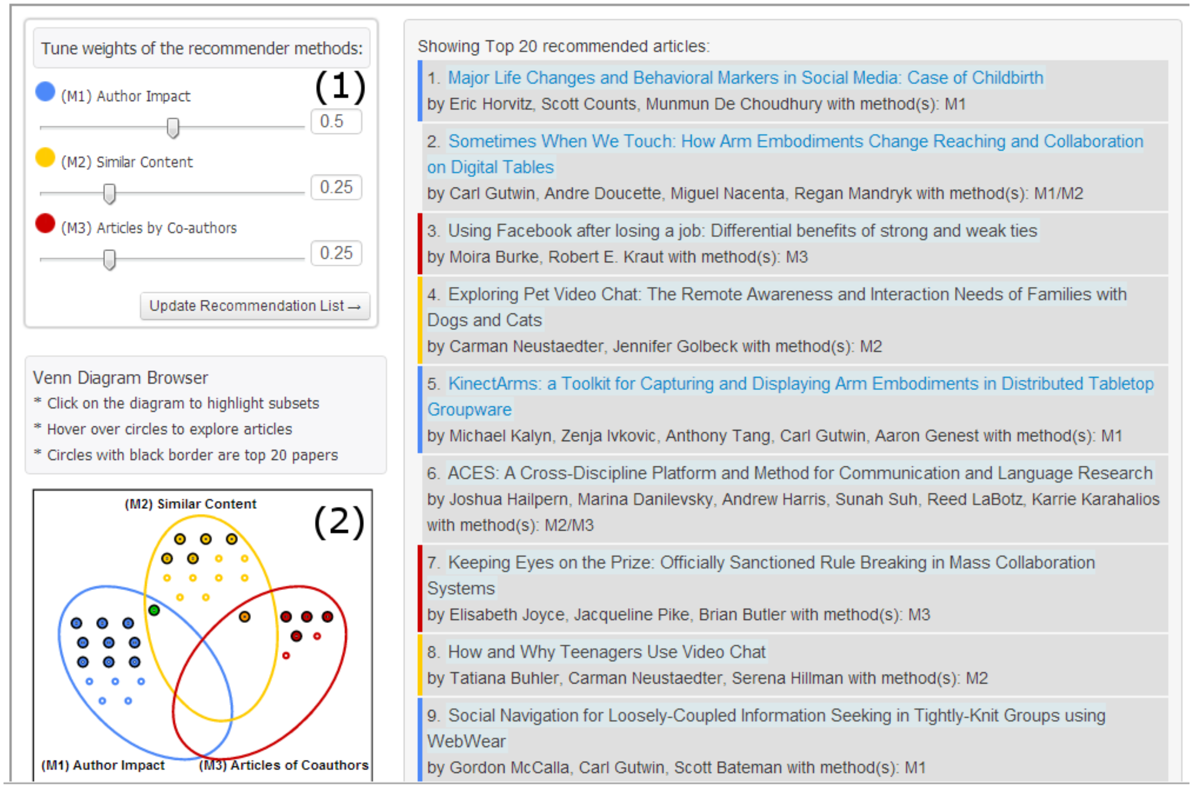
1.2.4 RQ4. Do user characteristics affect the influence of controllability on the user experience in a recommender system?

The influence of user characteristics on the user 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 already 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, as explained in 1.2.1 "RQ1. How does controllability affect user engagement with a recommender system?". For one thing, controllability and inspectability in a hybrid recommender are implemented in a different way in this research (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. Second, academic talks are very different from music bands or songs in their intrinsic characteristics and type of user consumption (Lamere, 2012), and therefore user characteristics might play a different role with respect to controllability and user experience with the system.

1.3 THE RECOMMENDER USER INTERFACES

The Studies 1 and 2 described in this dissertation contrast the user's response to two interfaces: controllable and non-controllable (baseline condition). In order to explain what kind of controllability is contrasted with the baseline condition, both interfaces are introduced here and their interactions are explained in more detail in chapter 3.0 Research Platform.

Figure 1 presents the controllable interface, where the user is able to control and inspect the recommendations. The component labeled with "(1)" allows the user to control the weight of each recommender method, from 0 to 1, by moving the sliders. The label "(2)" in Figure 1 indicates a Venn diagram that allows the user to explore the intersections among the different recommendation methods, i.e., which items have been recommended by either one, two, or three of the methods described. Finally, Figure 1 also presents the list of recommendations where the user can read the title, the list of authors, and a color to the left of each item indicates the method that produced that recommendation.



Tune weights of the recommender methods:

(1)

● (M1) Author Impact 0.5

● (M2) Similar Content 0.25

● (M3) Articles by Co-authors 0.25

Update Recommendation List →

Venn Diagram Browser

* Click on the diagram to highlight subsets

* Hover over circles to explore articles

* Circles with black border are top 20 papers

Showing Top 20 recommended articles:

1. Major Life Changes and Behavioral Markers in Social Media: Case of Childbirth
by Eric Horvitz, Scott Counts, Munmun De Choudhury with method(s): M1

2. Sometimes When We Touch: How Arm Embodiments Change Reaching and Collaboration on Digital Tables
by Carl Gutwin, Andre Doucette, Miguel Nacenta, Regan Mandryk with method(s): M1/M2

3. Using Facebook after losing a job: Differential benefits of strong and weak ties
by Moira Burke, Robert E. Kraut with method(s): M3

4. Exploring Pet Video Chat: The Remote Awareness and Interaction Needs of Families with Dogs and Cats
by Carman Neustaedter, Jennifer Golbeck with method(s): M2

5. KinectArms: a Toolkit for Capturing and Displaying Arm Embodiments in Distributed Tabletop Groupware
by Michael Kalyn, Zenja Ivkovic, Anthony Tang, Carl Gutwin, Aaron Genest with method(s): M1

6. ACES: A Cross-Discipline Platform and Method for Communication and Language Research
by Joshua Hailpern, Marina Danilevsky, Andrew Harris, Sunah Suh, Reed LaBotz, Karrie Karahalios with method(s): M2/M3

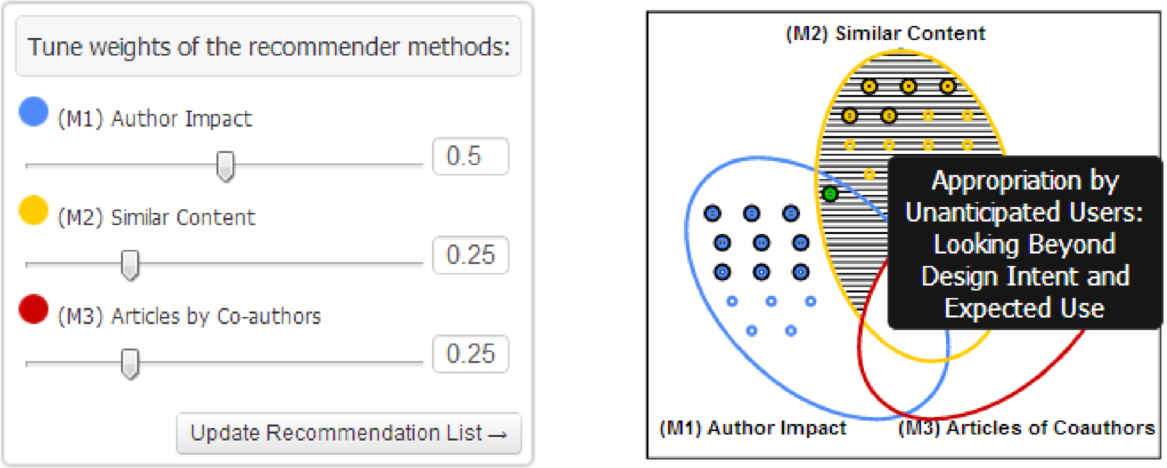
7. Keeping Eyes on the Prize: Officially Sanctioned Rule Breaking in Mass Collaboration Systems
by Elisabeth Joyce, Jacqueline Pike, Brian Butler with method(s): M3

8. How and Why Teenagers Use Video Chat
by Tatiana Buhler, Carman Neustaedter, Serena Hillman with method(s): M2

9. Social Navigation for Loosely-Coupled Information Seeking in Tightly-Knit Groups using WebWear
by Gordon McCalla, Carl Gutwin, Scott Bateman with method(s): M1

Figure 1. Screenshot of interactive controllable interface. In addition to browsing a list the articles, the user can control the list using (1) and (2), and inspect the items using (2).

Figure 2 presents a detailed view of the widgets of the controllable interface. Figure 2 a) shows three sliders, one for each of recommendation method: Author Impact, Similar Content, and articles written by co-authors. Figure 2 b) presents a Venn diagram, where each ellipse represents a recommendation method, with their colors matching those in the sliders interface. The circles inside the ellipses represent items recommended by each method, but only those shown in the recommendation list are highlighted with a black border. The image also shows that when the mouse icon hovers over the little circles, a tooltip showing the description of the talk is displayed.



Tune weights of the recommender methods:

(M1) Author Impact 0.5

(M2) Similar Content 0.25

(M3) Articles by Co-authors 0.25

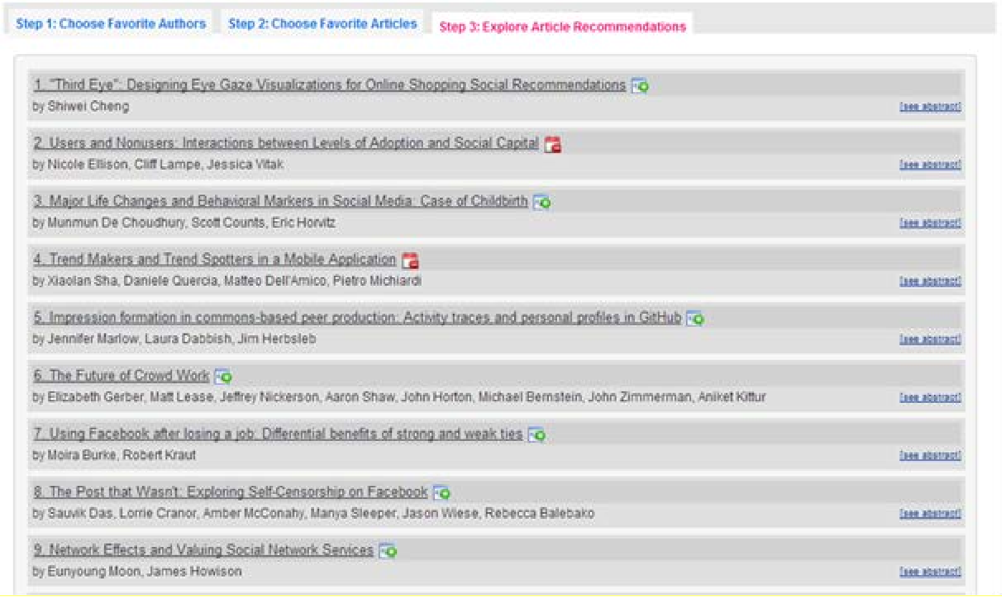
Update Recommendation List →

(M2) Similar Content
Appropriation by Unanticipated Users: Looking Beyond Design Intent and Expected Use
(M1) Author Impact
(M3) Articles of Coauthors

a) b)

Figure 2. Detailed view of a) the control widget that allows manipulating the weight of each method, and b) the Venn Diagram that allows users exploring the recommendations and the intersection between methods.

Finally, Figure 3 shows the non-controllable interface, where the user can explore a static list of recommendations ranked by relevance.


Step 1: Choose Favorite Authors
Step 2: Choose Favorite Articles
Step 3: Explore Article Recommendations

1. "Third Eye": Designing Eye Gaze Visualizations for Online Shopping Social Recommendations
by Shiwei Cheng [see abstract]

2. Users and Nonusers: Interactions between Levels of Adoption and Social Capital
by Nicole Ellison, Cliff Lampe, Jessica Vitak [see abstract]

3. Major Life Changes and Behavioral Markers in Social Media: Case of Childbirth
by Munmun De Choudhury, Scott Counts, Eric Horvitz [see abstract]

4. Trend Makers and Trend Spotters in a Mobile Application
by Xiaolan Sha, Daniele Quercia, Matteo Dell'Amico, Pietro Michiardi [see abstract]

5. Impression formation in commons-based peer production: Activity traces and personal profiles in GitHub
by Jennifer Marlow, Laura Dabbish, Jim Herbsleb [see abstract]

6. The Future of Crowd Work
by Elizabeth Gerber, Matt Lease, Jeffrey Nickerson, Aaron Shaw, John Horton, Michael Bernstein, John Zimmerman, Aniket Kittur [see abstract]

7. Using Facebook after losing a job: Differential benefits of strong and weak ties
by Moira Burke, Robert Kraut [see abstract]

8. The Post that Wasn't: Exploring Self-Censorship on Facebook
by Sauvik Das, Lorrie Cranor, Amber McConahy, Manya Sleeper, Jason Wiese, Rebecca Balebako [see abstract]

9. Network Effects and Valuing Social Network Services
by Eunyoung Moon, James Howison [see abstract]

Figure 3. Screenshot of the non-controllable condition. The interface presents a static list of conference talks, which is built using a hybrid recommendation method in background.

Each recommended talk is described by its important attributes: talk title, the authors, and the abstract, which is available after the user clicks on the link "see abstract." In addition, users can add the talk to their schedule by clicking on the icon at the right of the title.

1.4 CONTRIBUTIONS

Although controllability, inspectability and enriched visual interfaces for recommendations have been already studied in the past, this dissertation provides- the following contributions to the current literature:

1.4.1 Explainability based on different depths of field

There are many alternatives to visually explain recommendations to users, such as an ego network layout (O'Donovan et al., 2008) a circular layout connecting different users' interests (Gou, You, Guo, Wu, & Zhang, 2011), or parallel lists with links connecting recommendations with different sources (Bostandjiev et al., 2012). Inspired by our own previous work, which shows that presenting recommended items in clusters under several contexts of relevance increases the users' likelihood to choose those items (Verbert et al., 2013), I implemented a Venn diagram visualization that allowed users to see the big picture without losing the details, i. e., which items were recommended by one, two or more approaches, as shown in Figure 4, where each ellipse represents a recommendation method, and the little circles inside represent items recommended. The circles with a black border are presented in the final recommendation list. The others are less relevant but the user can still see if there are more recommended items in case she wants to explore.

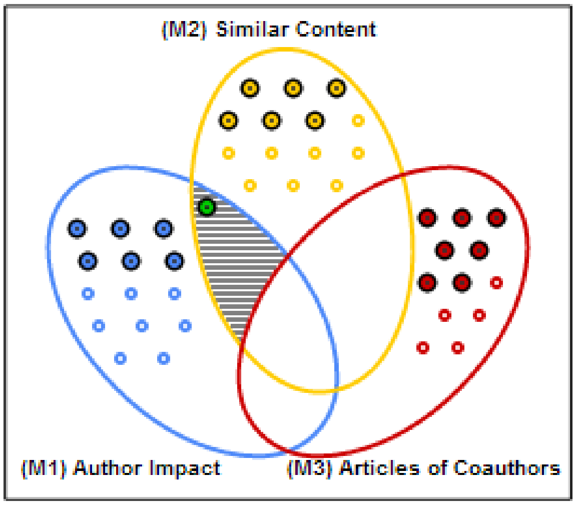


Figure 4. Venn diagram representing three recommender methods (ellipses M1, M2 and M3) and the items recommended by each method (circles). The shaded area highlights an item recommended by two methods.

The reason to pick this kind of visualization, beyond the obvious observation that it has not been used in previous research to present recommendations, is that it provides “different depths of field”. In (Lurie & Mason, 2007), the authors define the “depths of fields” of a visualization as “the extent to which it provides contextual overview versus detail information or enable decision makers to keep both levels in focus at the same time”. Using a 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). As already mentioned, in (Bostandjiev et al., 2012) the authors introduce a novel tool, TasteWeights, to control different sources of data and the weights influencing the final recommendation list, but their visualization doesn’t allow the subjects to visualize what they are missing from the combination of different neighbors or sources of data.

1.4.2 Controllability in a domain with different contexts of relevance

By the time of the writing this dissertation, three domains have been explored in studies over the influence of controllability and other variables on the user experience with a recommender system: energy-saving, music and jobs. They have particular characteristics that make it difficult to extrapolate their research results to a domain like recommendation of talks in conferences. For instance, Lamere mentions in (Lamere, 2012) that music has a very large item space compared to other domains (Apple's iTunes Catalog has 28 million songs[1]), the consumption time is low (average song's length is around 4 minutes), new technology has made music consumption highly interactive (user feedback includes playing, skipping, repeating, adjusting volume, etc.) and very high per item re-use. Although these are important differences from conference talks, both domains can provide different contexts of relevance to produce recommendations. This makes them suitable for implementing a hybrid recommender, which consequently supports studying user controllability on fusing recommendations from different relevance contexts. For instance, talks can be regarded as relevant based on the similarity of their content to previous user's bookmarks, based on the number of citations of their authors in a specific area, or simply based on the interest shown by users to attend to the talk (talk popularity), information available in Conference Navigator. The potential in this domain (talk recommendation) of using algorithms that yield different results underpins this research by allowing an investigation of whether allowing users to manually control the mixture of recommendations, i.e. allowing the users to control a hybrid recommender, will improve their user experience with the system.

1.4.3 Behavioral analysis

This is the first study on controllability in recommender systems that performs a behavioral analysis to measure the user experience in a controllable recommender interface. No previous study of controllability or explainability has looked deeper into behavioral measures, i.e., the actual user interaction with the interface to explain the influence of the aforementioned variables over the user experience. In (Attfield et al., 2011), the authors encourage the evaluation of user

[1] http://en.wikipedia.org/wiki/ITunes_Store accessed on February 11th 2013.

engagement using both surveys and objective measures, since the former rely on users' subjective, post-hoc interpretation, and their susceptibility to the halo effect, a "cognitive bias whereby the perception of one trait is influenced by the positive evaluation of another trait of a person or object", as described by the authors.

1.5 LIMITATIONS AND DELIMITATIONS

System. Conference Navigator is a web system developed at the PAWS lab that supports conference attendees. The studies that are described in the next subsections were conducted using this system for the particular domain of talk recommendation. This may affect the generalizability of the results.

What is Being Recommended and to Whom. The system recommended conference talks. As talks I considered long and short research papers in the main conference, but also keynotes, posters and workshop papers. The potential users of this recommender were conference attendees, but also people who don't necessarily attend, i. e., people who use CN3 and might be interested in the content of a particular conference.

Comparing Objective Metrics Between Controllable and Non-Controllable Interface. Is important to consider that in the case of the non-controllable condition, each user is associated with a single metric value (e.g. the precision over a single list of recommendations) and the means tested in the hypotheses are means over the set of users. On the other side, a single subject under the controllable condition has multiple measures of accuracy and rank due to the interactive nature of the interface. Each time the user changed the weights of the sliders or clicked on the ellipses of the Venn diagram, a new list was presented and a new metric -

precision, MRR, or nDCG- was calculated. Using precision as an example, for a specific user a single measure of precision under the non-controllable condition is compared to the average of "several precisions" under the controllable interface, calculated over each recommendation list that the user created during the task.

Devices Used to Access the Application. As a web system, Conference Navigator can be accessed through any modern web browser, and nowadays browsers are available in many different kind of devices. The analysis conducted in this dissertation does not differentiate whether the users accessed the system via a desktop computer, a laptop, a mobile device or a tablet. Although visits to Conference Navigator through a handheld browser are a small percentage of the total (2.5% between May 1st and July 31 of 2013), this important dimension is not controlled in this dissertation and should be taken into consideration in future studies.

The *Target* of Recommendations. In Study 2 users were asked to bookmark relevant papers from different versions of the iConference. Since the diversity of topics in this conference might match few relevant items (talks) to some of the subjects of this study, we tried to increase the likelihood of users' picking more items by asking them to bookmark at least 15 conference talks relevant not only to themselves, but also to other colleagues. This setting adds complexity to the analysis since it introduces an additional dimension in the evaluation (for whom is the user considering a specific talk relevant), but for this step was necessary to ensure a minimum number of items required to evaluate objective and subjective metrics. From a more pragmatic perspective, this represents a common scenario: sometimes people discover talks in a conference that are relevant to colleagues, and they might be likely to share them.

1.6 DEFINITION OF TERMS

Clustermap: This is a type of visualization that shows clusters of items and connections between entities connected to these clusters. It was introduced as Aduna (Klerkx & Duval, 2009) to visualize social tagging system, where the entities are users, URLs, and user-tags, and the items inside the clusters are the bookmarks.

Conference Navigator: This is an online conference support system[2] developed by the PAWS lab at the University of Pittsburgh that allows conference attendees to explore the talks in a conference, schedule a personalized program of talks, or receive talk recommendations, among many other features.

Controllability: This is a property of a system that allows the user to control, intervene or take action over the recommendation process. There are many alternatives to allow the users controlling the recommendation process, from things as simple as ratings, to changing the weight of specific features or attributes of the items, such as in (Hijikata et al., 2012)

Depths-of-field: The concept of "depths-of-fields" of visualization has been defined as "the extent to which it provides contextual overview versus detail information or enable decision makers to keep both levels in focus at the same time" (Lurie & Mason, 2007).

Explainability: A property or capacity of a system of being explained by a user. For instance, if the user receives a list of recommended talks and the user understands why she was recommended that list, we can say that the system offers good explainability.

Inspectability: By inspectability, I mean "the capacity of the system of being inspected and explainable to the user" (Knijnenburg, Bostandjiev, et al., 2012).

[2] http://halley.exp.sis.pitt.edu/cn3/

Recommender System: Though there are several definitions, I select this one by (McNee et al., 2006a) for its simplicity and clarity: “a recommender or recommendation system aims to help a user or a group of users in a system to select items from a crowded item or information space”.

2.0 BACKGROUND WORK

This chapter describes past work that supports and motivates this investigation. First, work performed by other researchers is described in the section titled *Related Work*, with a focus on five areas: controllability in recommender systems, talk recommendation in events, transparency and explainability, user-centric evaluation of recommender systems, and hybrid recommenders. The second section of this chapter, called *Preliminary Work*, describes the results of my previous research on exploratory visual recommendation that supports some of the research questions addressed in this dissertation.

2.1 RELATED WORK

2.1.1 Control, Inspectability and User Intervention in Recommender Systems

The positive effect of increased user control in online shopping (Ariely, 2000) and user interaction in decision making (Lurie & Mason, 2007) has been studied for more than a decade. In adaptive systems, Sherman and Shortlife (1993) introduced a user-adaptable interface called PODIUM that allowed physicians to control the creation of clinical user interfaces with positive results. Only recently has user-control been methodically investigated in the context of recommender systems. Knijnenburg et al. (2011) studied the effect of different interaction mechanisms on an energy-saving recommender system. . They conclude that the best interaction mechanisms depends on user characteristics; for instance, expert users (with more domain knowledge) reported higher user satisfaction with interfaces that provided more control compared to the novices users, who would be more satisfied with an interface that provides the recommendation without many controllable variables. Bostandjiev et al. (2012) introduce a visual hybrid interactive music recommender called TasteWeights and they perform a small-scale study (N=32) 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 the users have "tuned" the importance of different data sources and neighbors using the visualization, was better with bigger interaction and explainability (the full interface), same with the general user experience. Using the same TasteWeights framework, but only considering social recommendation (Facebook contacts of the center user) of music, Knijnenburg et al. (2012) perform a user study on the influence of control and inspectability (N=267) 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 of friends produced slightly better recommendation than controlling the weight of items, but the effects of both are additive so providing both in a real setting is recommended. Hijikata et al. (2012) also explore control in the context of music recommendation. They explore four different ways to let users intervene the recommendation process: ratings, context, content attribute and user profile edition. With a user study (N=84), they show 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 find 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.

To sum up, this previous literature on user controllability (Hijikata et al., 2012; Knijnenburg, Bostandjiev, et al., 2012; Knijnenburg et al., 2011) shows a positive relationship between user satisfaction and user control, with moderating factors such as user expertise, engagement with the domain and level of participation. However, these studies investigate controllability understood or implemented as:

1. Letting users sort a list of items and customize the importance (weights) of items' attributes, in an attribute-based system that recommends energy-saving measures. (Knijnenburg et al., 2011) study 5 different user interaction methods (TopN, Sort, Explicit, Implicit, and Hybrid), their interaction with personal characteristics (domain knowledge, trusting propensity and persistence) and their effect on several variables such us perceived control, understandability, user interface satisfaction, perceived system effectiveness and choice satisfaction, among others. This study concludes that in order to provide the best user experience, the design of attribute-based recommender systems should consider interaction methods depending on different user characteristics.

2. Letting the users update their preferences (music bands), the importance of different data sources, and the importance of the k-nearest neighbors from each data source in a social music recommender system (Bostandjiev et al., 2012). This article introduces the TasteWeights system, a hybrid visual recommender system. The authors also present a user study with N=32 users showing that users rated with higher utility the recommendation lists produced after they inspected and tuned the weights of data sources and neighbors.

3. Giving the users control over the recommendation generation by letting them intervene either by ratings, users' context, items' coarse grained attributes or items' fine grained attributes in a music recommender system. In this study, (Hijikata et al., 2012) investigate the effect of different levels of control (they called them *levels of user intervention*) on user satisfaction. Though they do a deeper data exploration than the previous two studies, finding a relation between user interest, recommender precision and user satisfaction with levels of user interest, their statistical analysis is weaker and provides only preliminary evidence of these effects, rather than a clear conclusion.

Compared to these studies, this dissertation contributes to the human factors dimension of recommender systems by investigating 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 user's satisfaction.

2.1.2 Talk Recommendation in Events

Although recommendation of documents and research articles has been explored in several studies (McNee et al., 2002), (Denis Parra & Brusilovsky, 2009) and (Ekstrand et al., 2010), recommendation of talks in a conference setting or in other events is a more specific task, and has not been explored so extensively. (Brusilovsky, Parra, Sahebi, & Wongchokprasitti, 2010) study the effects of combining diverse sources of information to recommend talks in a conference system. They concluded that using additional user information from another system such as citeulike (an online reference manager), or the tags provided by users, in addition to users' explicit preferences for talks in the system, had a positive effect on the quality of the recommendations. In a follow-up study, (Sahebi, Wongchokprasitti, & Brusilovsky, 2010) study combining different sources of user preference to provide recommendations of open lectures at University of Pittsburgh and Carnegie Mellon University in the CoMeT system, and find similar results. (Minkov, Charrow, Ledlie, Teller, & Jaakkola, 2010) also study future event recommendation by comparing a content-based recommender using RankSVM with a proposed collaborative filtering method they call LowRank. After running two user studies at MIT and CMU talk series over 15 weeks, they conclude that LowRank performs better than other state-of-the-art matrix factorization techniques. All the evaluations in the papers discussed in this paragraph relied exclusively on accuracy-based evaluation, and no user-centric evaluation of satisfaction, explainability or trust in the system is reported.

2.1.3 Transparency and Explainability in Recommender Systems

(Herlocker, Konstan, & Riedl, 2000) introduce the idea of explaining recommendations as a mean to make the system more transparent to users' decision 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 prove it an call 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 design and evaluation of explanations in recommender systems (Tintarev & Masthoff, 2011).

2.1.4 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 this metrics, recent research has increasingly considered user-centric evaluation measures such as perceived diversity, controllability and explainability. For instance, (Ziegler, McNee, Konstan, & Lausen, 2005) studied the effect of diversification in lists of recommended items, (Tintarev & Masthoff, 2007) investigated on recommender systems' transparency, (Cramer et al., 2008) studied explainability in recommender systems, and (Knijnenburg, Bostandjiev, 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 address this issue. On one side, (Pu et al., 2011) propose 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, Willemsen, et al., 2012) define dimensions and relations between them (objective systems aspects, subjective system aspects, experience, interaction, situational characteristics and personal characteristics), but encourage the user of this framework to choose her own constructs based on some specified guidelines.

2.1.5 Hybrid recommenders: Fusing recommendations of several methods

In the section 3.0 “Research Platform”, I explain the recommendation methods used in these studies. In the studies described in the upcoming sections the user will see a single list of recommendations, which requires fusing the results of the three different recommendation methods. This is a common task in hybrid recommender systems, but there are several fusion alternatives. In (Burke, 2002), the author describes seven methods, which are presented in Table 1. The most straightforward alternative for the *algorithm-blending* approach I propose in this research is the *mixed hybridization* method. Considering this is the same one used by Bostandjiev et al. to evaluate *TasteWeights* (Bostandjiev et al., 2012), I decided to choose this method to make the procedure and results comparable.

Table 1. Hybridization methods surveyed by (Burke, 2002)

Hybridization method	Description
Weighted	The scores (or votes) of several recommendation techniques are combined together to produce a single recommendation.
Switching	The system switches between recommendation techniques depending on the current situation.
Mixed	Recommendations from several different recommenders are presented at the same time
Feature combination	Features from different recommendation data sources are thrown together into a single recommendation algorithm
Cascade	One recommender refines the recommendations given by another.
Feature augmentation	Output from one technique is used as an input feature to another.
Meta-level	The model learned by one recommender is used as input to another.

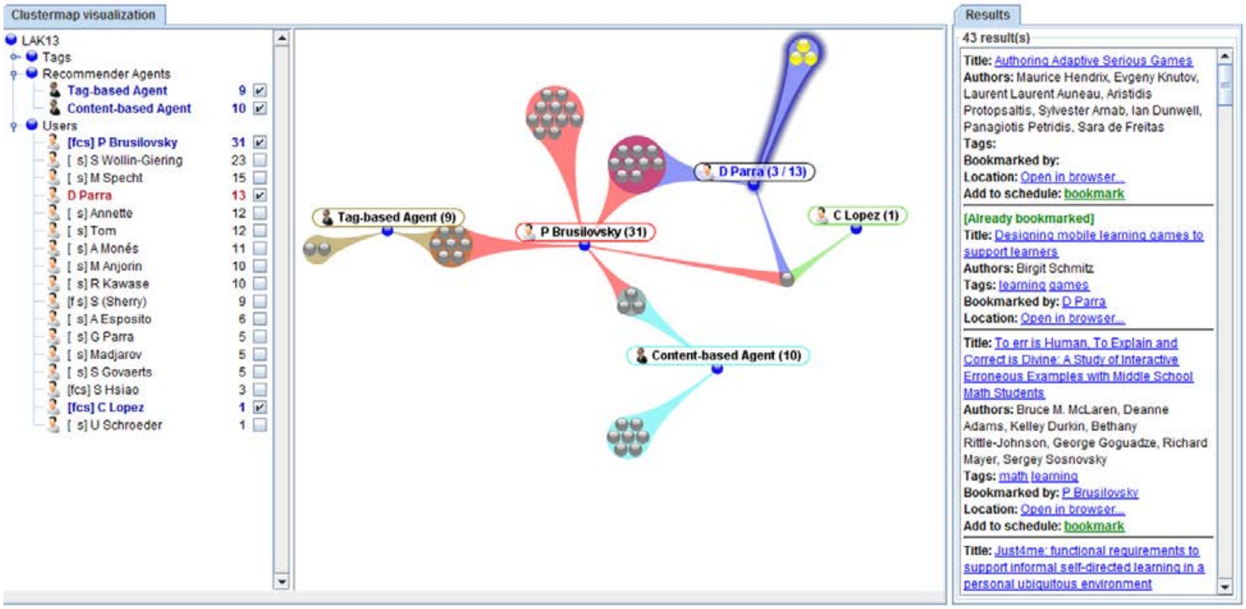
Taking into account that the relevance scores returned by the three methods are different ranges (in once case score go from 0 to 1, in another case from 1 to 10, and the normalized scores does not seem to be comparable), the final score of each recommended item will be based on its rank in the recommendation list of each method, and with more importance given to items recommended by more than one method. The fusion will be performed in such a way that the score of a recommended item *src(rec$_i$)* will be given by:

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

Where *M* is the set of all methods available to fuse –in this proposal: author impact, content-based, and co-author neighborhood-, $rank_{reci,mj}$ 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 $|_{Mreci|}$ represents the number of methods by which item rec_i was recommended.

2.2 PRELIMINARY WORK: TALKEXPLORER

In (Verbert et al., 2013), we embedded into Conference Navigator a visual interactive tool called Aduna[3] that we had adapted to allow users to explore talks in a conference from multiple perspectives of relevance-talks bookmarked by users, suggestions of recommender agents and talks marked with specific tags.



LAK13
- Tags
- Recommender Agents
  - Tag-based Agent 9 ☑
  - Content-based Agent 10 ☑
- Users
  - [fcs] P Brusilovsky 31 ☑
  - [s] S Wollin-Giering 23 ☐
  - [s] M Specht 15 ☐
  - D Parra 13 ☑
  - [s] Annette 12 ☐
  - [s] Tom 12 ☐
  - [s] A Monés 11 ☐
  - [s] M Anjorin 10 ☐
  - [s] R Kawase 10 ☐
  - [f s] S (Sherry) 9 ☐
  - [s] A Esposito 6 ☐
  - [s] G Parra 5 ☐
  - [s] Madjarov 5 ☐
  - [s] S Govaerts 5 ☐
  - [fcs] S Hsiao 3 ☐
  - [fcs] C Lopez 1 ☑
  - [s] U Schroeder 1 ☐

43 result(s)

Title: Authoring Adaptive Serious Games
Authors: Maurice Hendrix, Evgeny Knutov, Laurent Laurent Auneau, Aristidis Protopsaltis, Sylvester Arnab, Ian Dunwell, Panagiotis Petridis, Sara de Freitas
Tags:
Bookmarked by:
Location: Open in browser...
Add to schedule: bookmark

[Already bookmarked]
Title: Designing mobile learning games to support learners
Authors: Birgit Schmitz
Tags: learning games
Bookmarked by: D Parra
Location: Open in browser...

Title: To err is Human, To Explain and Correct is Divine: A Study of Interactive Erroneous Examples with Middle School Math Students
Authors: Bruce M. McLaren, Deanne Adams, Kelley Durkin, Bethany Rittle-Johnson, George Goguadze, Richard Mayer, Sergey Sosnovsky
Tags: math learning
Bookmarked by: P Brusilovsky
Location: Open in browser...
Add to schedule: bookmark

Title: Just4me: functional requirements to support informal self-directed learning in a personal ubiquitous environment

Figure 5. Screenshot of Aduna TalkExplorer embedded in Conference Navigator

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

Figure 5 shows a screenshot of our integration into Conference Navigator (TalkExplorer) displaying three main panels: entity selection panel (on the left), canvas panel (at the center) and the talk description panel (on the right). We conducted a think-aloud study during 2 conferences (Hypertext and UMAP 2012), where we asked users to bookmarks talks to select and explore 3 different entities (users, recommender agents, and tags).

Figure 6 shows clusters of talks associated to entities with intersections of talks. In the figure, the grey and yellow circles represent papers –people could click on them and add them to their conference schedule-, and the blue circles with labels represent people, agents or tags. The figure shows three main entities, the user "D Parra" (with 13 bookmarks), the "Tag-based Agent" (with 10 papers to recommend) and the "Content-based agent" (with 10 papers to recommend).

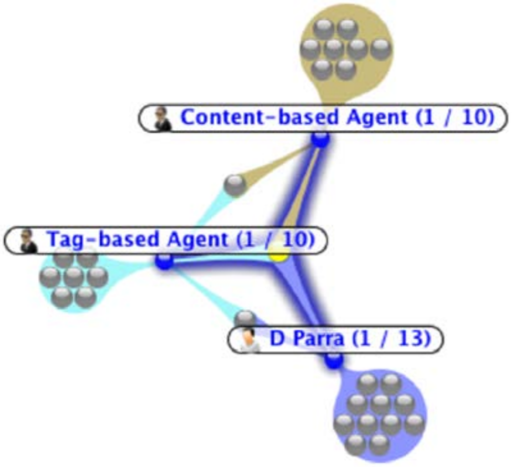


Figure 6. Screenshot of clustermap at an intersection of talks between three entities.

In Figure 6, the user "D Parra" has a cluster of 11 talks (represented by 11 grey circles clustered in a blue wrapper, at the bottom-right corner of the image), and 2 talks shared with the recommender agents, one represented by a yellow circle right in the middle of image –a talk shared with the 2 recommender agents- and a grey one –shared only with the tag-based agent.

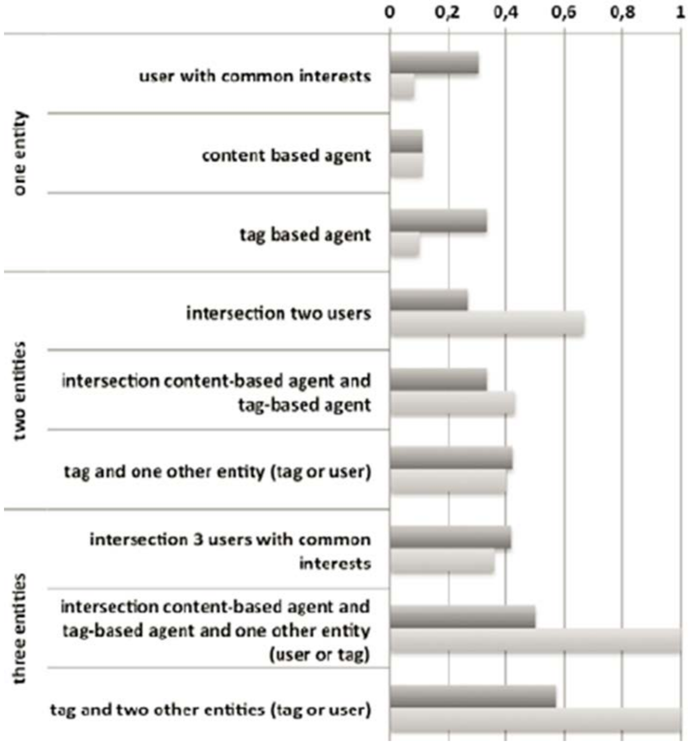


Figure 7. Results of effectiveness (dark grey) and yield (light grey) in the TalkExplorer study.

Similarly, the “Tag-Based agent” has 7 talks wrapped in a light-blue cluster and 3 other talks shared with the other entities, whereas the “Content-based Agent” has 8 talks to Similarly, the “Tag-Based agent” has 7 talks wrapped in a light-blue cluster and 3 other talks shared with the other entities, whereas the “Content-based Agent” has 8 recommended talks wrapped into a brown cluster and two other talks shared. The results of our think-aloud study revealed that users often explore clusters of talks that show relationships between entities (as the one shown in Figure 6), and the probability of selecting an item increases when the items is at the intersection of several entities’ clusters. We measured this behavior with two metrics: yield and efficiency, as can be seen in Figure 7 that shows results from our TalkExplorer study (Verbert et al., 2013).The first one is analogous to precision for one cluster of talks; yield is the ratio between the number of talks selected and the number of talks available in that cluster. On the other hand, efficiency measures the ratio between the number of clusters in which the user picked a talk and the number of clusters that the user explored.

These results are interesting, but, due to the protocol of the task, there are limitations in generalizing them: we asked users to interact with at least three entities, which might be not the case in a real setting. On the other hand, a more detailed analysis might unveil more subtle differences in the interactions among specific entities. Finally, although the intersection of talks in the visualization was effective, the Aduna clustermap visualization might not be as natural as a Venn diagram for showing intersections between sets of entities. Investigating this assumption motivates the implementation of a Venn diagram to let users filter intersections of talks among different recommender methods.

3.0 RESEARCH PLATFORM

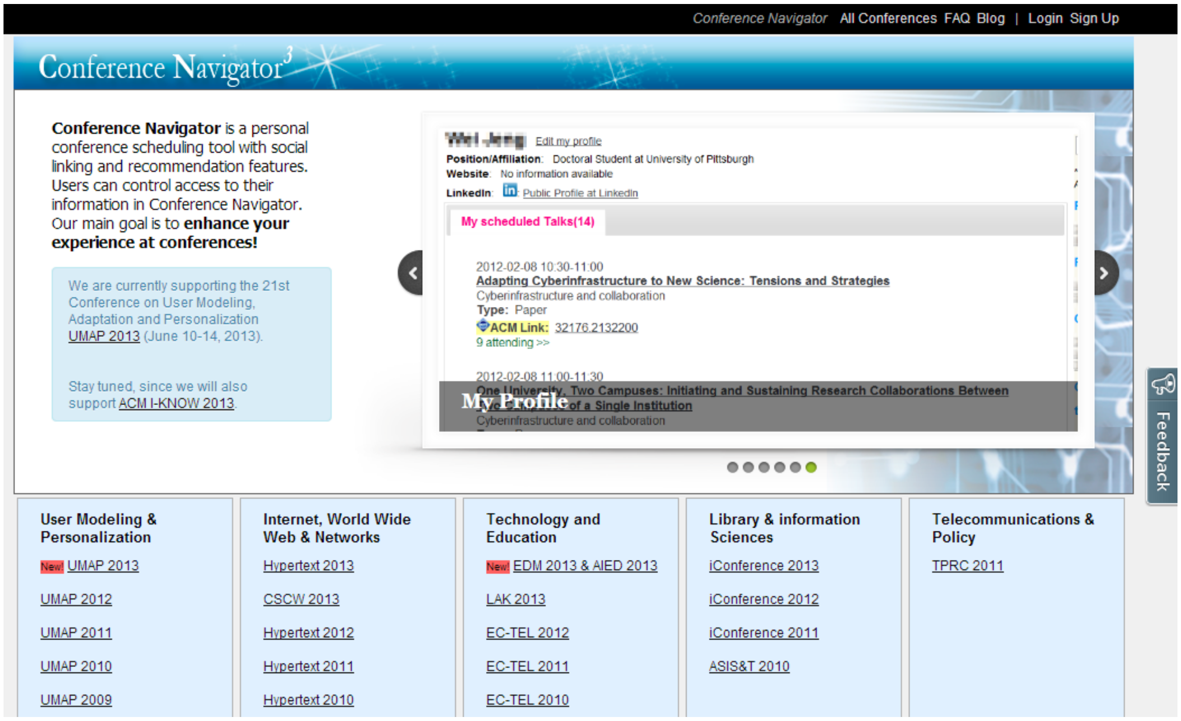
In this section I describe the four main components of the research platform used to conduct the user studies in this dissertation: the Conference Navigator system, the recommender interfaces (controllable and non-controllable), the overall system architecture, and the recommendation approaches used to generate talk recommendations.

3.1 CONFERENCE NAVIGATOR

In order to conduct the two user studies used to address the research questions, an innovative controllable interface was implemented in an existing online conference support system. In the next two subsections I introduce the system, Conference Navigator, and then provide a general description of the recommender interface used for this investigation.

3.1.1 Introducing Conference Navigator

Conference Navigator 3 (Denis Parra, Jeng, Brusilovsky, López, & Sahebi, 2012) is the third version of a system developed by the PAWS Lab at the University of Pittsburgh (Farzan & Brusilovsky, 2008). Its objective is to support conference attendees by providing several useful tools, like a program and proceedings schedule, enhanced with social navigation features like a list of popular bookmarked talks and paper recommendations.



Conference Navigator All Conferences FAQ Blog | Login Sign Up

Conference Navigator3

Conference Navigator is a personal conference scheduling tool with social linking and recommendation features. Users can control access to their information in Conference Navigator. Our main goal is to **enhance your experience at conferences!**

We are currently supporting the 21st Conference on User Modeling, Adaptation and Personalization UMAP 2013 (June 10-14, 2013).

Stay tuned, since we will also support ACM I-KNOW 2013.

[redacted] Edit my profile
Position/Affiliation: Doctoral Student at University of Pittsburgh
Website: No information available
LinkedIn: Public Profile at LinkedIn

My scheduled Talks(14)

2012-02-08 10:30-11:00
Adapting Cyberinfrastructure to New Science: Tensions and Strategies
Cyberinfrastructure and collaboration
Type: Paper
ACM Link: 32176.2132200
9 attending >>

2012-02-08 11:00-11:30
One University, Two Campuses: Initiating and Sustaining Research Collaborations Between Two ... of a Single Institution
Cyberinfrastructure and collaboration

My Profile

Feedback

User Modeling & Personalization	Internet, World Wide Web & Networks	Technology and Education	Library & information Sciences	Telecommunications & Policy
New! UMAP 2013	Hypertext 2013	New! EDM 2013 & AIED 2013	iConference 2013	TPRC 2011
UMAP 2012	CSCW 2013	LAK 2013	iConference 2012	
UMAP 2011	Hypertext 2012	EC-TEL 2012	iConference 2011	
UMAP 2010	Hypertext 2011	EC-TEL 2011	ASIS&T 2010	
UMAP 2009	Hypertext 2010	EC-TEL 2010		

Figure 8. Screenshot of the homepage of Conference Navigator, the system used to conduct the field and laboratory studies.

Figure 8 shows a screenshot of the homepage of the system. By the writing of this dissertation, Conference Navigator has been used to support over 24 conferences since 2008, with a focus on conferences like UMAP (User Modelling, Adaptation and Personalization), Hypertext, ECTEL and the iConference. The decision to use Conference Navigator to conduct both the field and laboratory studies described in this dissertation was based on three main factors. For a field study, a certain minimum number of user interacting with the system is required, and Conference Navigator has shown increasing user activity, as seen in Table 2. This made using Conference Navigator possible for the field study in CSCW 2013. Another important factor to consider is the amount of past information stored by the system to produce recommendations. Conference Navigator stores usage information about each conference, such as user bookmarks, ratings, and other user actions that allows for the production of, for instance, non-personalized popularity-based recommendations for any conference.

Table 2. Usage statistics of CN 2 and 3 over different conferences. Number of users who bookmarked/tagged talks and connected to other users in relation to the total number of attendees and availability of recommendations for talks (T) and people (P)

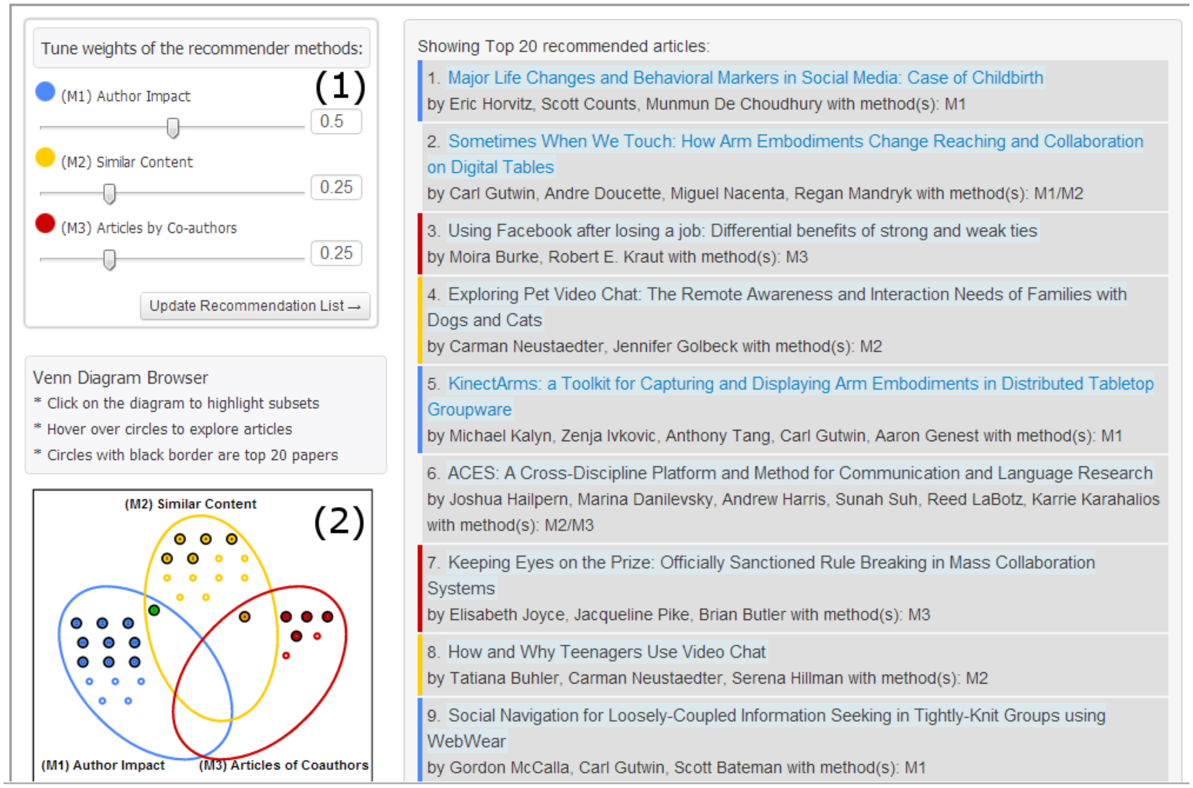
	Conference Navigator 2				Conference Navigator 3			
	UMAP 2009	HT 2009	UMAP 2010	ASIST 2010	iConf 2011	ECTEL 2011	HT 2011	UMAP 2011
Recommended			T		T	T	T, P	T, P
Attended	171	141	134	550	474	141	111	176
Bookmarked	24	18	30	58	109	36	42	69
Tagged	12	13	12	11	2	21	18	21
Connected	-	-	-	-	2	8	3	8
Bookmarks	177	114	266	471	1327	416	499	1019
%Bookmarked	14.04%	12.77%	22.39%	10.55%	23.00%	25.53%	37.84%	39.20%
%Tagged	7.02%	9.22%	8.96%	2.00%	0.42%	14.89%	16.22%	11.93%

Finally, Conference Navigator allows the tracking of different users’ actions (clicks on different visual widgets, bookmarking actions etc.), which can be helpful in a laboratory study to answer research questions by controlling for additional factors. With regard to the benefits to the current implementation of Conference Navigator, implementing a new talk recommender is expected to increase the system’s usage. Although paper recommendation is one of the system’s main features, it has not attracted too much attention from users compared to other features like top items, conference schedules and proceedings. An important outcome from the implementation of a visual controllable recommender is attracting more people to use this feature and, therefore, benefit from it.

3.2 RECOMMENDER INTERFACES

3.2.1 Details of the User Interface Interactions

The two recommender interfaces, controllable and non-controllable, were already introduced in section 0 “The Recommender User Interfaces”. Figure 9 presents all the components of the experimental controllable interface together.



Tune weights of the recommender methods:

● (M1) Author Impact 0.5

● (M2) Similar Content 0.25

● (M3) Articles by Co-authors 0.25

Update Recommendation List ⟶

Venn Diagram Browser

* Click on the diagram to highlight subsets

* Hover over circles to explore articles

* Circles with black border are top 20 papers

Showing Top 20 recommended articles:

1. Major Life Changes and Behavioral Markers in Social Media: Case of Childbirth
by Eric Horvitz, Scott Counts, Munmun De Choudhury with method(s): M1

2. Sometimes When We Touch: How Arm Embodiments Change Reaching and Collaboration on Digital Tables
by Carl Gutwin, Andre Doucette, Miguel Nacenta, Regan Mandryk with method(s): M1/M2

3. Using Facebook after losing a job: Differential benefits of strong and weak ties
by Moira Burke, Robert E. Kraut with method(s): M3

4. Exploring Pet Video Chat: The Remote Awareness and Interaction Needs of Families with Dogs and Cats
by Carman Neustaedter, Jennifer Golbeck with method(s): M2

5. KinectArms: a Toolkit for Capturing and Displaying Arm Embodiments in Distributed Tabletop Groupware
by Michael Kalyn, Zenja Ivkovic, Anthony Tang, Carl Gutwin, Aaron Genest with method(s): M1

6. ACES: A Cross-Discipline Platform and Method for Communication and Language Research
by Joshua Hailpern, Marina Danilevsky, Andrew Harris, Sunah Suh, Reed LaBotz, Karrie Karahalios with method(s): M2/M3

7. Keeping Eyes on the Prize: Officially Sanctioned Rule Breaking in Mass Collaboration Systems
by Elisabeth Joyce, Jacqueline Pike, Brian Butler with method(s): M3

8. How and Why Teenagers Use Video Chat
by Tatiana Buhler, Carman Neustaedter, Serena Hillman with method(s): M2

9. Social Navigation for Loosely-Coupled Information Seeking in Tightly-Knit Groups using WebWear
by Gordon McCalla, Carl Gutwin, Scott Bateman with method(s): M1

Figure 9. Screenshot of the experimental interface, the visual controllable recommender.

In this section I present a detailed description of the interface actions that users were able to perform when interacting with the visual controllable recommender. The following subsections describe these actions in three main visual components: the list of recommended items, the sliders widget and the Venn diagram widget.

3.2.1.1 List of recommended items

The list of recommended items is very similar under both conditions–controllable and non-controllable interface-so the interaction described here applies to both interfaces, with the exception of hovering the mouse over the color bar located at the left side of each recommended item. This color bar hints at the method or methods used to recommend the associated paper, which is an explanation feature only available in the controllable interface. Figure 10 shows a screenshot of the recommended list with the actions available, which were:

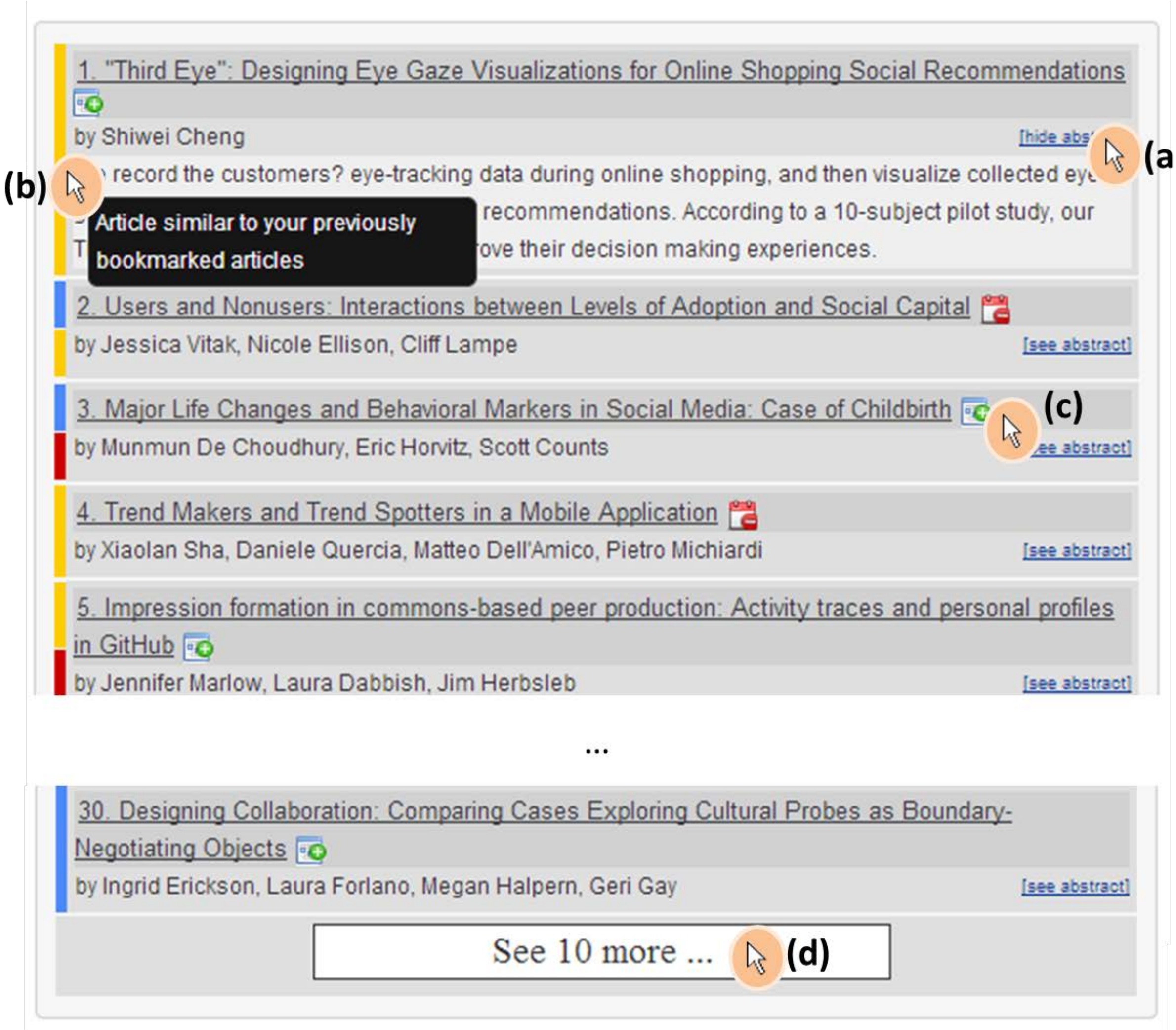
a) Open and Close Abstract: by clicking on the link provided by each paper title, the users could see the abstract of the article.

b) Hover over color bar: only available in the controllable condition, users could hover over the color bar to obtain an explanation of the method used to recommend the paper.

c) Bookmark a paper: at the very end of each paper's title, an icon indicates if the paper is bookmarked or not. This same icon allows the user to bookmark or remove the paper from the relevant items.

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."

e) Rate a paper: Although not shown in Figure 10, in the CSCW study there was a rating bar below the author list showing five numbers (1-2-3-4-5) that allowed the user to rate the relevance of a paper in a scale from 1 (not relevant at all) to 5 (strongly relevant).



1. "Third Eye": Designing Eye Gaze Visualizations for Online Shopping Social Recommendations

by Shiwei Cheng

… record the customers? eye-tracking data during online shopping, and then visualize collected eye … recommendations. According to a 10-subject pilot study, our … ove their decision making experiences.

Article similar to your previously bookmarked articles

2. Users and Nonusers: Interactions between Levels of Adoption and Social Capital

by Jessica Vitak, Nicole Ellison, Cliff Lampe

3. Major Life Changes and Behavioral Markers in Social Media: Case of Childbirth

by Munmun De Choudhury, Eric Horvitz, Scott Counts

4. Trend Makers and Trend Spotters in a Mobile Application

by Xiaolan Sha, Daniele Quercia, Matteo Dell'Amico, Pietro Michiardi

5. Impression formation in commons-based peer production: Activity traces and personal profiles in GitHub

by Jennifer Marlow, Laura Dabbish, Jim Herbsleb

...

30. Designing Collaboration: Comparing Cases Exploring Cultural Probes as Boundary-Negotiating Objects

by Ingrid Erickson, Laura Forlano, Megan Halpern, Geri Gay

Figure 10. Screenshot of the recommended items list. The arrows highlight the actions that the user could perform in the recommender interface.

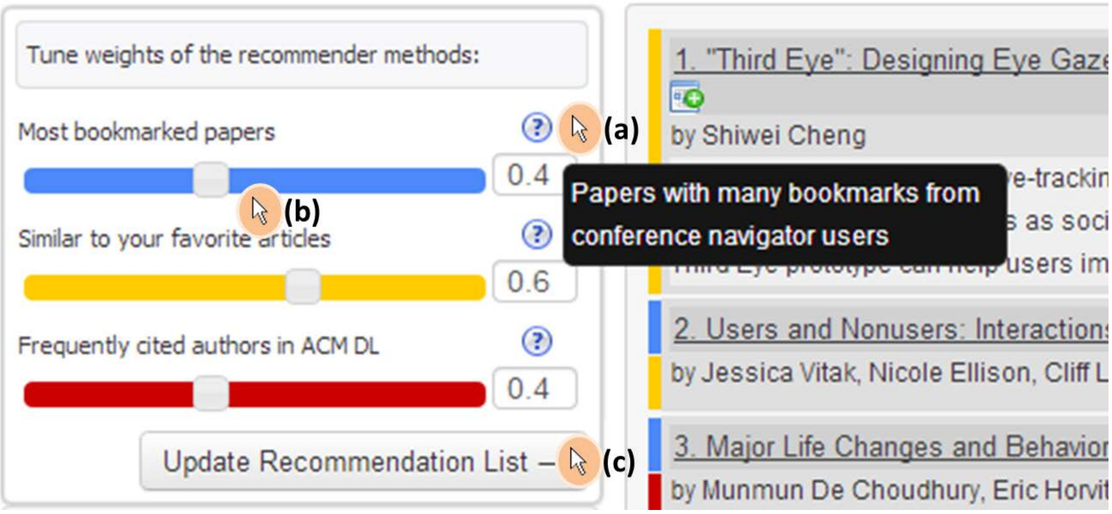
3.2.1.2 Sliders

The sliders are only present in the controllable interface. They allow the user to re-sort the list of recommended items based on the importance that they assign to each method, represented by different colors as shown in Figure 11. The interactions available on this widget are:

a) Hover over explanation icon: this action allows the user to obtain a more detailed explanation of the method by displaying a black floating dialog.

b) Move sliders: by moving the sliders, the users change the relative importance of each method used to generate the list. Users can also input a weight directly in the text box besides each slider.

c) Update recommendation list: after moving the sliders to adjust the importance of each method, the user must click on the button "Update Recommendation List" in order to sort the list of recommendations on the right-side panel.



Tune weights of the recommender methods:

Most bookmarked papers: 0.4

Similar to your favorite articles: 0.6

Frequently cited authors in ACM DL: 0.4

Papers with many bookmarks from conference navigator users

(a)
(b)
(c)
Update Recommendation List
1. "Third Eye": Designing Eye Gaze
by Shiwei Cheng
e-trackin
s as soci
users im
2. Users and Nonusers: Interaction
by Jessica Vitak, Nicole Ellison, Cliff L
3. Major Life Changes and Behavior
by Munmun De Choudhury, Eric Horvit

Figure 11. Screenshot of the sliders widget. The arrows highlight the actions that the user could perform in the controllable interface.

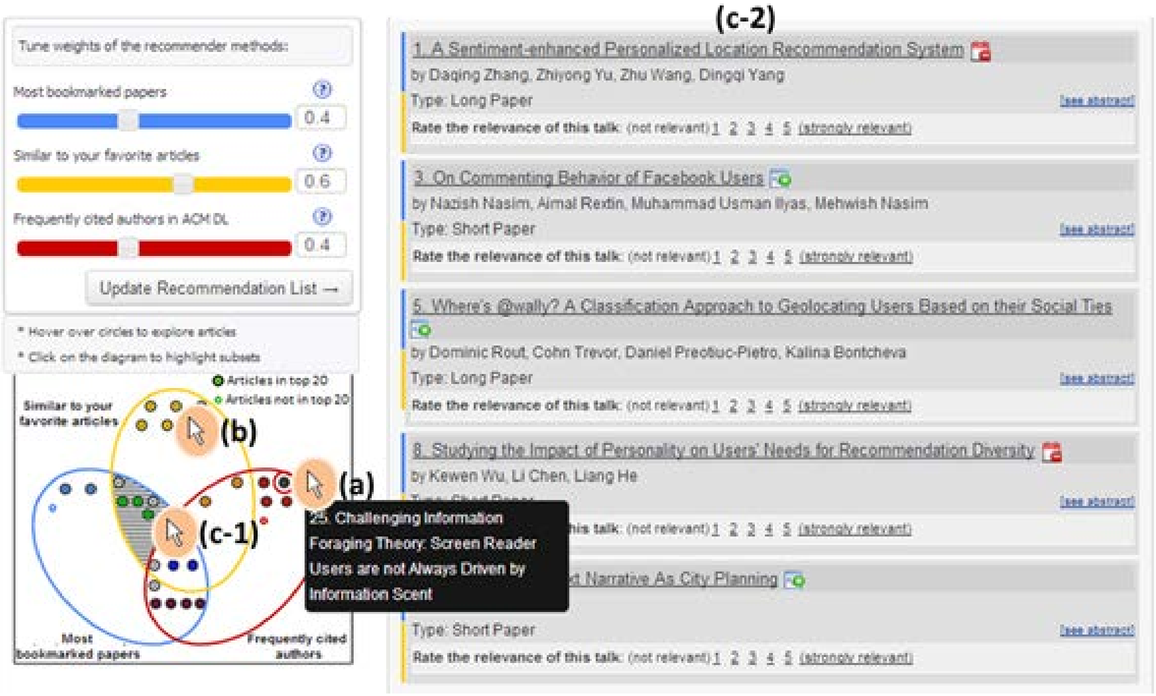
3.2.1.3 Venn diagram

The Venn diagram, shown in Figure 12, is only present in the controllable interface. It provides to the users a high-level view of the items recommended by each method and the "intersections" between them, i.e., the papers recommended by more than one method. The interactions available on this widget are:

a) Hover over the circle: Each circle represents a talk. This action opens a small floating dialog with the title of the talk being explored.

b) Click on a circle: with this action, the system scrolls up or down to that paper in the list on the right-side panel.

c) Clicking on a Venn diagram area (ellipse): When the user clicks on an area or sub-area of the Venn diagram (c-1 in Figure 12), this action allows the user to filter the list on the right panel (c-2), showing only those articles that were recommended by the method or methods represented by the ellipse or intersection of ellipses. The filtering behavior differed from studies 1 to 2. In the case of Study 1, the talks selected stayed visible in their same position in the list and the rest of talks were hidden, leaving empty spaces visible between talks. Unlike Study 1, in Study 2 the list was collapsed as shown in Figure 12, so that there is no distance between items 1 and 3.



Tune weights of the recommender methods:

Most bookmarked papers 0.4

Similar to your favorite articles 0.6

Frequently cited authors in ACM DL 0.4

Update Recommendation List →

* Hover over circles to explore articles
* Click on the diagram to highlight subsets

25. Challenging Information Foraging Theory: Screen Reader Users are not Always Driven by Information Scent

1. A Sentiment-enhanced Personalized Location Recommendation System
by Daqing Zhang, Zhiyong Yu, Zhu Wang, Dingqi Yang
Type: Long Paper [see abstract]
Rate the relevance of this talk: (not relevant) 1 2 3 4 5 (strongly relevant)

3. On Commenting Behavior of Facebook Users
by Nazish Nasim, Aimal Rextin, Muhammad Usman Ilyas, Mehwish Nasim
Type: Short Paper [see abstract]
Rate the relevance of this talk: (not relevant) 1 2 3 4 5 (strongly relevant)

5. Where's @wally? A Classification Approach to Geolocating Users Based on their Social Ties
by Dominic Rout, Cohn Trevor, Daniel Preotiuc-Pietro, Kalina Bontcheva
Type: Long Paper [see abstract]
Rate the relevance of this talk: (not relevant) 1 2 3 4 5 (strongly relevant)

8. Studying the Impact of Personality on Users' Needs for Recommendation Diversity
by Kewen Wu, Li Chen, Liang He
[see abstract]
his talk: (not relevant) 1 2 3 4 5 (strongly relevant)

t Narrative As City Planning
Type: Short Paper [see abstract]
Rate the relevance of this talk: (not relevant) 1 2 3 4 5 (strongly relevant)

Figure 12. Screenshot that highlights the actions available on the Venn diagram widget.

3.2.1.4 Summary of Actions Available on Each Interface

Table 3 presents a summary of the actions available on the controllable and the non-controllable interfaces. There are 6 actions in common, the rest are available only on the controllable interface, since they are related to the sliders and to the Venn diagram visual widgets.

Table 3. List of actions tracked in the recommender interfaces (Controllable and non-controllable).

Action	Description	Visual Widget	Controllable,	Non-Controllable
clickRetrieveList	Retrieve initial list of recommendations	Recommender interface	X	X
scheduling	Bookmark a talk	Talk	X	X
unscheduling	Remove bookmark	Talk	X	X
seeMore	Expand list or recommendations	Recommender List	X	X
clickOpenAbstract	Open abstract of talk	Talk	X	X
clickCloseAbstract	Close abstract of talk	Talk	X	X
changeSlider[N]	Change weight of method N	Slider Widget	X	
clickUpdateList	Update recommendation list	Slider Widget	X	
hoverMethod[N]Explain	Show explanation of method N	Slider Widget	X	
hoverCircle[N]	Mouse over circle (talk) on the subarea (method) N	Venn diagram	X	
clickEllipse[N]	Click Ellipse (Venn diagram) to filter list by method N	Venn diagram, list of talks	X	

3.3 RECOMMENDATION APPROACHES

The main source of user control in this research proposal is implemented by letting users combine different recommendation algorithms. In order to make these algorithms work, different sources of information and user preference feedback are needed, depending on the approach, but, recommending talks for a specific event is particularly challenging compared to traditional movie or music recommendation. First, the cold-start problem is evident, since the talks to be presented are new to the research community, and probably very few people have had the chance to read and provide implicit (web page visits) or explicit feedback (ratings or bookmarks) before the conference starts. Second, there is a very short period of time for the system to capture feedback from enough users interacting with enough items in order to produce recommendations. Depending on the conference, the list of talks with their respective program schedule can be available to public as early as a few months to a few weeks before the conference. Finally, during the event itself, the recommendations are only useful before the talks have taken place; afterwards, they become irrelevant. Given these considerations, I used information crawled from the ACM library to fast start (Brusilovsky et al., 2010) users' and items' profiles, which will allow me to produce recommendations for the popularity, content-based and collaborative filtering methods.

Two different sets of recommendation approaches were implemented. The first set of three methods was implemented for study 1, conducted during the CSCW 2013 conference. Given different conditions for the dataset, a different set of three methods was used for study 2, conducted as a controlled laboratory study using iConference data. In the following subsections, I describe these two sets of methods and how they were combined into a hybrid recommender to produce the final lists of recommendations.

3.3.1 Methods Used in the CSCW 2013 study

3.3.1.1 Popularity Based on Author Impact

To rank talks by their popularity, given that the usual starting condition is of a lack of feedback from users, I calculated the popularity of the papers based on the popularity of their authors in terms of the numbers of citations they had received. This means that popularity was understood as *expected impact popularity*. To calculate the popularity of papers' authors based on how frequently cited they have been in the past, I used a dataset crawled from the ACM Digital Library. From there, the procedure to obtain the popularity of each paper is the following:

a) List the papers for each conference where the study was held (iConference 2013, CSCW 2013, HT 2013, UMAP 2013)

b) Obtain the authors 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, I mean a function that gives a relevance score to a paper based on the maximum "number of references" among the authors of that paper.

3.3.1.2 Content-Based Algorithm

Generating recommendations with the content-based algorithm has two scenarios:

a) The subject already has a user account in Conference Navigator and she has provided feedback or ratings to papers in previous related conferences. In this case, I can use the title and abstract of those papers as a "user model" and find similar papers in the current conference (iConference 2013, CSCW 2013, HT 2013, and UMAP 2013) to produce the recommendations.

b) The subject is not a previous user of Conference Navigator, or is a previous user but has not provided feedback yet. In this case, the user was shown a list of the authors of the conference, and then asked to pick three of them. The interface then presented some papers (from previous conferences) from these authors and asked the user to select the most relevant papers from that list. With this feedback, the content-based recommender created a user model and matched papers from the current conference whose content was similar. The information needed (past papers from the conference authors) was taken from the ACM digital library dataset.

For the two scenarios previously described, 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 are calculated by TF*IDF (term frequency * inverse document frequency), as explained in (Manning, Raghavan, & Schtze, 2008). To find relevant documents, the matching is performed by using the Lucene[4] function MoreLikeThis[5], 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.

3.3.1.3 Co-author Neighborhood Inspired by Collaborative Filtering

Initially, I thought of using collaborative filtering as one of the alternative recommendation algorithms. However, this algorithm is the most affected by the cold-start. In order to produce collaborative filtering recommendations, either user-based or item-based, not only must the

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

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

subject provide enough feedback (which can be elicited during the study), but there is also a need for feedback from enough users, and feedback given to enough items in the dataset. There are two possible solutions: one that allows collaborative filtering using feedback provided by users in Conference Navigator, but under the risk that the user-talk preference matrix is too sparse, and another that requires implementing an alternative algorithm inspired by collaborative filtering. My choice was the second one, involving a method more similar to a spreading activation (Troussov, Parra, & Brusilovsky, 2009) strategy, which uses the co-authorship network. We start by some initial nodes (authors) chosen by the users as seeds, then we spread the activation (relevance) through the co-authors of the authors chosen in the first step, finally to choose talk recommendations from papers in the current conference written by these co-authors.

To summarize, the steps to produce the recommendations list were the following:

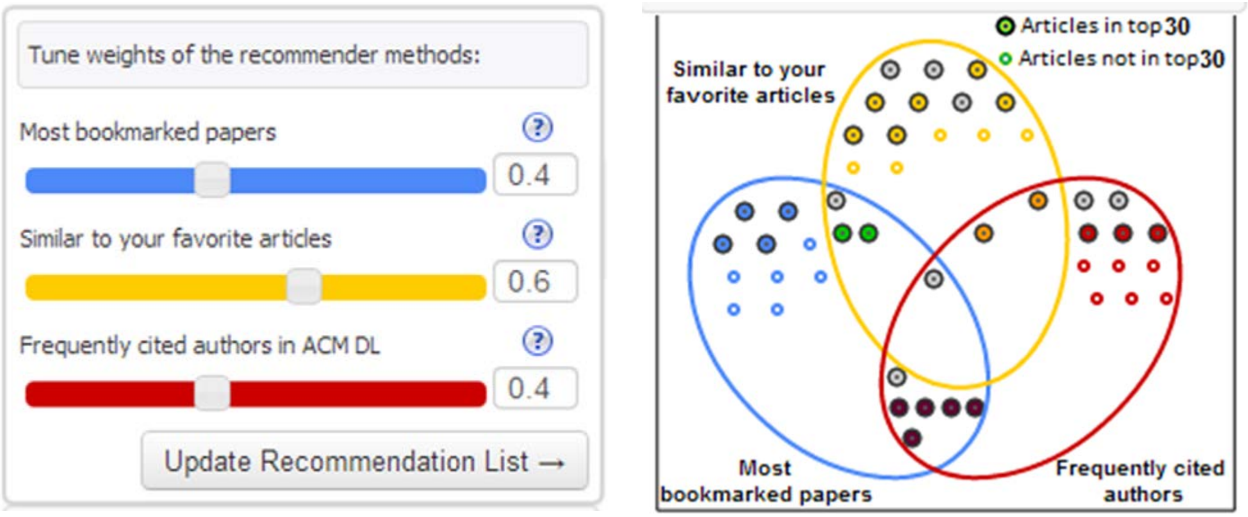
a) Ask subjects to select their most relevant authors from the list of conference authors. The users' neighborhood will be composed of those authors and their coauthors.

b) Then, the potential set of recommended items are papers in the current conference (CSCW 2013) authored by the co-authors of the researchers chosen by the center user in step a).

c) The items to recommend were ranked based on the number of relevant co-authors that wrote a certain paper, where the weight of each author was proportional to the number of references that the author had in the ACM DL database, i.e., the expected impact of the co-author.

3.3.2 Methods Used in the iConference Study, Hypertext 2013 and UMAP 2013

Unlike CSCW 2013, in the subsequent studies one of the methods used to produce the recommendation list was changed. The content-based and “popularity based on author impact” methods remained the same, but the “co-author neighborhood” method described in section 3.3.1.3 was removed and replaced by recommending articles based on their number of bookmarks, i. e, the number of times that actual Conference Navigator users added them to their personal schedule. The reason for this change is the nature of the conference. CSCW has a long history (25 years at the moment of this writing) and there is a good chance that the attendees at this conference: a) identify authors that they selected in the recommender interface, and b) that co-authors of the favorite authors are publishing in the current conference.



Tune weights of the recommender methods:

Most bookmarked papers: 0.4

Similar to your favorite articles: 0.6

Frequently cited authors in ACM DL: 0.4

Figure 13. Screenshot shows the methods used to produce recommendations in the studies 2 and 3.

The iConference does not have these characteristics. It was only indexed in the ACM DL dataset for 2 years (2011 and 2012), and the Conference Navigator only has data available for 3 years of the conference (2011 to 2013). The diverse content of the conference makes it less likely to find co-authors of famous authors publishing in in it, reducing the potential for recommendations. Figure 13 shows the final methods presented in the visual controllable recommender interface for studies 2 (CSCW) and 3 (HT and UMAP): Most bookmarked papers (bookmarking popularity), similarity to your favorite articles (content-based recommender), and frequently cited in ACM library (popularity based on author impact). Note that the content-based recommender is the only personalized recommendation method.

3.4 SYSTEM ARCHITECTURE

The main part of the infrastructure used in this research is Conference Navigator, which is located in the server halley.exp.sis.pitt.edu (from now on, *halley*). The content-based recommender, however, is in the columbus.exp.sis.pitt.edu server (hereinafter, *columbus*). There were two reasons to implement the content-based recommender on a different server: a) to avoid affecting the performance of Conference Navigator, and b) to avoid conflicts with existing services in halley.

Figure 14 summarizes the architecture of the system. Conference Navigator is a web system written in PHP 5.0 that stores the data in a MySQL 5.1.8 database server. The main file that implements the recommendation interface (both controllable and non-controllable) is hybrid_recsys.php. Most of the information displayed by this page in the client's browser comes directly from halley's database. However, the user interaction in the controllable and non-controllable recommender makes extensive use of JavaScript for the three visual components of the page: the sliders, the Venn diagram, and the list of recommendations displayed in the right panel.

The JavaScript library Jquery 1.7.2 is used for the filtering interactions in the sliders, and in the recommendation list. It is also used to make the AJAX calls to the columbus server to retrieve the personalized content-based recommendations. The columbus server stores the conference papers in a Lucene[6] index, and Solr[7] works as a bridge between the AJAX calls from hybrid_recsys.php and the recommendations generated through the Lucene API. The interaction of the Venn diagram is implemented mainly with the Javascript library Raphael 2.0[8],

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

[7] http://lucene.apache.org/solr/

[8] http://raphaeljs.com/

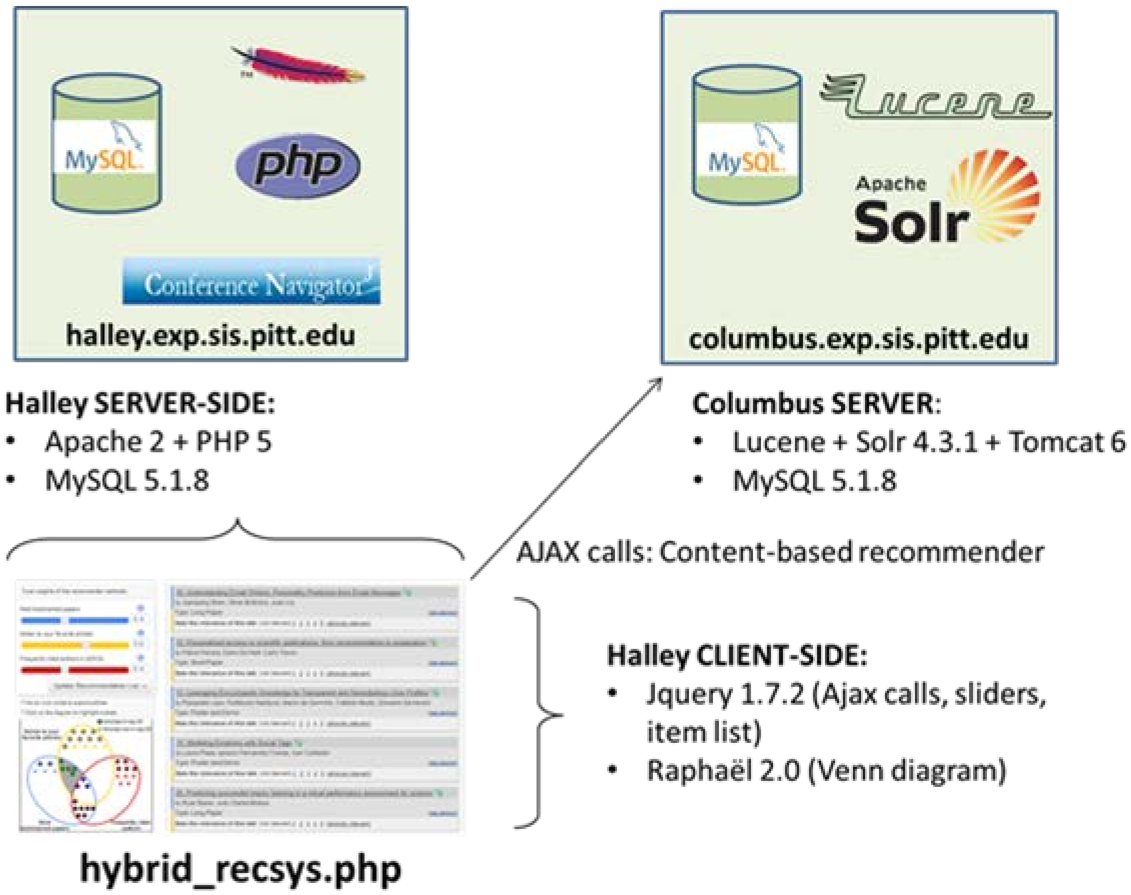


Figure 14. Diagram that summarizes the system architecture used to implement the recommender interface in Conference Navigator.

4.0 MEASURES

This chapter describes all the measures used to evaluate the hypotheses on engagement and user experience, as well as the user characteristics expected to have an effect on user behavior with the recommendation interfaces. Studies 1 and 2 had different settings, so some of the evaluation measures were used in both studies and others were used in only one. The last section of this chapter summarizes which measures and user characteristics were considered in each study.

The subjective measures to evaluate user engagement have been adapted from a framework that comprises four dimensions: focused attention, perceived usability, endurability and novelty (O'Brien & Toms, 2010). Moreover, the measures related to user experience with the recommender interfaces are adapted from (Pu et al., 2011) and from (Knijnenburg, Willemsen, et al., 2012).

4.1 EVALUATING USER ENGAGEMENT

One of the research goals of this dissertation is to measure effect of controllability on the user engagement with the system, I will evaluate this variable with subjective and objective measures based on the guidelines presented in (O'Brien & Toms, 2010) and (Attfield, Kazai, Lalmas, & Piwowarski, 2011).

- Subjective Measures (***M1***):): These are measures captured by questionnaires. Many of these factors or constructs (groups of questions) will be suitable for evaluating the recommender user experience in general as well. O'Brien and Tom's identify 6 final constructs that play a role in evaluating user engagement, I considered the four that were most relevant for evaluating user engagement:

    - Focused attention: The concentration of mental activity; concentrating on one stimulus only and ignoring all others.

      I lost track of time while I was using the recommender interface

    - Perceived usability: Users' perceived effort in using the system, their ability to accomplish their tasks, the navigation and organization of the system.

      I became familiar with the recommender interface very quickly.

    - Endurability: the likelihood to remember things that we have enjoyed and a desire to repeat a fun activity.

      I would use this recommender system again for another conference in the future.

    - Novelty: This factor is composed of questions that "… spoke to the curiosity evoked by or participants' interest in the shopping task."

      When looking at the list of recommended talks I am interested to examine which recommendation method has been used.

Objective Measures (***M2***):

- Individual behavioral measures
    - Number of talks explored using the recommender interface
    - Number of talks bookmarked using the recommender interface
    - Number of clicks in the recommender interface
    - Amount of time in the recommender interface
    - Number of visits after using the recommender interface.

4.2 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 elaborate user-centric evaluation frameworks for recommender systems have been proposed. ResQue, introduced by (Pu et al., 2011), and one introduced by (Knijnenburg, Willemsen, et al., 2012). The first one, ResQue, describes several variables (they call them constructs, since each one can be measured by more than one question in a questionnaire) grouped into four high level layers: *perceived system qualities*, *users' beliefs* (as a result of system qualities), *subjective attitudes*, and their *behavioral intentions*. The second framework, introduced by (Knijnenburg, Willemsen, et al., 2012), distinguishes several dimensions but not the specific variables or questions that should be included in each one. Their framework consists of *objective system aspects, subjective system aspects*, *user experience*, *interaction*, and *personal and situational characteristics*.

In the following lists, a perceived system quality is followed by a question that was asked in some of the surveys used in the studies.

4.2.1 Related to Perceived System Qualities

- Explanation: The recommender explains why the items are recommended to me
- Interaction Adequacy: I find it easy to provide my preferences (ratings) to the system
- Recommendation Accuracy: The items recommended to me matched my interests.
- Recommendation Diversity: The items recommended to me are diverse
- Information Sufficiency: The information provided is sufficient to make a decision
- Interface Adequacy: the interface (labels and layout) is clear and adequate

4.2.2 Related to User Beliefs

- Transparency: I understood why the items were recommended to me
- Control: The recommender allows me to control the initial set of recommendations
- Perceived Usefulness: The recommender helped me to find the ideal items
- Perceived Ease of Use: I became familiar with the system very easily

4.2.3 Related to User Attitudes

- Trust and Confidence: The recommender made me confident of my selection/decision

- Overall Satisfaction: Overall, I am satisfied with the recommender

4.2.4 Related to Behavioral Intentions

- User Intention: I would suggest my colleagues to use this recommender system when they attend a conference in the future.

4.2.5 Objective metrics

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

- Average rating: I compare the conditions by calculating the mean over the average rating of each user under a particular condition. The mean average user rating of a condition *i* will be calculated as:

$$\bar{r_i} = \frac{\sum_{u \in U_i} \bar{r_u}}{|U_i|}$$

Where $\overline{r_i}$ represents the average rating under the condition *i, u* stands for user, U_i for the set of all users in condition *i,* and $\overline{r_u}$ represents the average rating of user *u* under condition *i*.

- Precision@k: This metric allow us to measure the accuracy of a list of *k* recommendations (Manning, Raghavan, & Schtze, 2008). For a given user *u*, the precision@k of a list of recommended items is

$$precision@k = \frac{|rel|}{k}$$

where |*rel*| is the number of relevant items recommended and *k* the number of items in the list. I can average this metric over the whole set of users in a condition to compare which condition presents a better accuracy based on this metric.

- 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). The average precision of the recommended list of items can be expressed as:

  $$AveP = \frac{\sum_{k=1}^{n}(P(k) \times rel(k))}{|relevant\ items|}$$

  Where *n* is the total number of items in the list, *P(k)* is the precision at cut point *k*, *rel(k)* is a function equal to 1 if the item is relevant and 0 if the item is not relevant, and |*relevant items*| is the number of relevant item in the list.

- nDCG: The name 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 an nDCG = 1. If we consider only two levels of relevancy (1: relevant item, 0: non-relevant item), then the nDCG of a list of recommendations is calculated as:

  $$\mathrm{nDCG} = \frac{DCG}{IDCG}$$

DCG is the Discounted Cumulative Gain of the list divided by the ideal Discounted Cumulative Gain (iDCG) when the list in ranked in the correct order (most relevant items at the top, less relevant and not relevant at the bottom). NDCG is defined as:

$$\text{DCG}= \sum_{i=1}^{k} \frac{2^{rel_i} -1}{\log_2(i+1)}$$

Where *k* is the number of items in the list, and *rel*$_i$ is the relevance score that the subject assessed over the recommended item *i*. *rel$_i$* can take the values 1 (relevant) or 0 (note relevant).

4.3 USER CHARACTERISTICS

Based on previous studies, these are personal and situational characteristics expected to interact with controllability. These variables are only considered in Study 2 (iConference study). They are:

- 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 the 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 SUMMARY OF MEASURES USED IN EACH STUDY

Table 4 summarizes the metrics used in each study, and it also lists the user characteristics considered in the analysis of Study 2.

Table 4. Measures and user characteristics considered on each user study.

	Study 1	Study 2		Studies 1 & 2
	Objective measures	Objective measures	User characteristics	Subjective measures
RQ1.User Engagement	Number of: talks explored, talks bookmarked, clicks, visits back, time spent	Number of: talks explored, talks bookmarked, time spent		focused attention, perceived usability, endurability & novelty
RQ2. User Experience	--	Avg user rating, precision@k, MAP, nDCG		perceived system qualities, user beliefs, user attitudes, & behavioral intentions
RQ3. User characteristics on engagement	--	Number of: actions, explored talks, time spent // number of actions on: sliders, Venn diagram	Age, gender, native speaker / Expertise, community status, trusting propensity, experience recsys	focused attention, perceived usability, endurability & novelty
RQ4. User characteristics on user experience	--	MAP		perceived system qualities, user beliefs, user attitudes, & behavioral intentions

5.0 STUDY 1

The first evaluation of the system was conducted with a between-subjects field user study during the CSCW 2013 conference, which was held in San Antonio, TX. Using the online conference support system Conference Navigator (Denis Parra et al., 2012) that allows users to access the conference program, proceedings, papers' details, and also to schedule their favorite talks, we provided the visual controllable recommender as a new feature.

5.1 DESIGN

5.1.1 Participants

Attendees of the conference that signed into Conference Navigator were randomly assigned to one of the two conditions (between-subjects design): either a controllable or a non-controllable interface. After signing into the system, users were redirected to a page with a promotional message, inviting them to try either the controllable or the non-controllable recommender. They were also given the opportunity to skip the message and continue using other features of Conference Navigator. Of a total of 161 people that viewed the promotional page, 37 used the recommender.

5.1.2 Hypotheses for Study 1

Based on the results of (Bostandjiev et al., 2012; Hijikata et al., 2012; Knijnenburg, Bostandjiev, et al., 2012), a controllable and inspectable interface is expected to improve user engagement and experience with a recommender system, depending on user characteristics. The hypotheses addressed in this study, which do not include user characteristics, are justified by the aforementioned studies, but their results may not be replicable with the new kind of control and inspectability introduced in this dissertation by the Venn diagram visualization. The following hypotheses were addressed in this study:

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

H1. Having a controllable interface will result in more engaged users, with the following objective measures increasing in a controllable interface:

1. Number of talks explored using the recommender interface
2. Number of talks bookmarked using the recommender interface
3. Number of clicks in the recommender interface
4. Amount of time spent in the recommender interface (minutes)
5. Number of visits after using the recommender interface

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

H2. A user in a controllable recommender interface will have a better user experience than a user in a non-controllable recommender interface, measured by mean user average rating.

H3. A user in a controllable recommender interface will have a better user experience than a user in a non-controllable recommender interface, measured by subjective metrics presented in a post-study survey. The dimensions evaluated in this survey that are expected to be more positive in the controllable interface are:

1. Perceived System Qualities
2. User Beliefs
3. User Attitudes
4. Behavioral Intentions

5.1.3 Study 1 Procedure

The normal workflow for this study requires four steps, as summarized in Table 5. At the end of the conference, the users of the system were sent an e-mail with a survey inquiring about their experience using the recommender system.

Table 5. Workflow of actions needed to get to the recommendation list in study 1.

Step 1	**Step 2**	**Step 3 (optional)**	**Step 4 (recommend list)**
In Conference Home Page -> Click on link to the Hybrid recommender Figure 16	Choose relevant authors from previous CSCW conferences Figure 18	Choose the papers from the authors of previous step more relevant to your interests (optional) Figure 19	List of recommendations, with (Figure 20) and without controllability (Figure 21)

Figure 15 shows graphically the procedure of Study 1, where the blocks under *User actions* represent the steps 2 to 4 in Table 5. The numbers below the boxes indicate the number of users participating actively on each step, reflecting the process' attrition. Each user that signed into Conference Navigator was randomly assigned to one of two conditions: the non-controllable recommender (a static list of recommendations) and the visual controllable recommender.

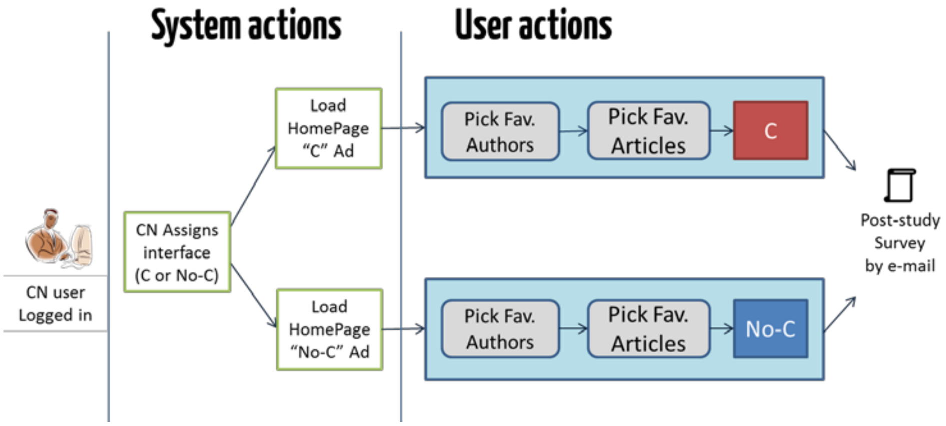
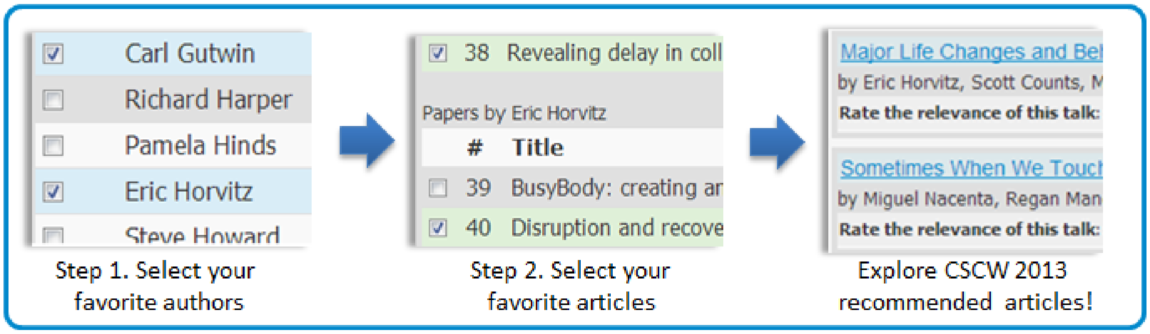


Figure 15. Between-subjects procedure of Study 1

Each condition had a different promotion message in the home page of the CSCW 2013 conference in Conference Navigator, as seen in Figure 16 and Figure 17.



☑ Carl Gutwin
□ Richard Harper
□ Pamela Hinds
☑ Eric Horvitz
□ Steve Howard

Step 1. Select your favorite authors

☑ 38 Revealing delay in coll

Papers by Eric Horvitz

# Title
□ 39 BusyBody: creating ar
☑ 40 Disruption and recove

Step 2. Select your favorite articles

Major Life Changes and Be
by Eric Horvitz, Scott Counts, M
Rate the relevance of this talk:

Sometimes When We Touc
by Miguel Nacenta, Regan Man
Rate the relevance of this talk:

Explore CSCW 2013 recommended articles!

Figure 16. Promotion in conference homepage in Conference Navigator, non-controllable interface.

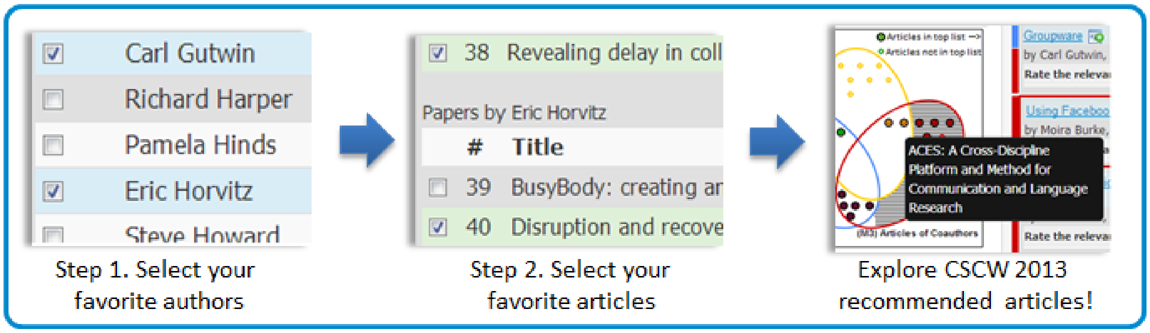
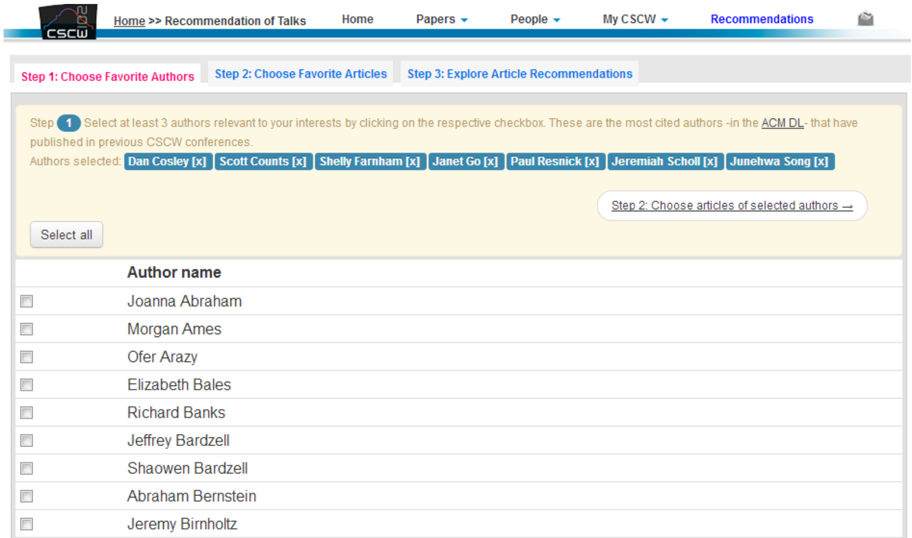

☑ Carl Gutwin
□ Richard Harper
□ Pamela Hinds
☑ Eric Horvitz
□ Steve Howard
Step 1. Select your favorite authors
☑ 38 Revealing delay in coll
Papers by Eric Horvitz
# Title
□ 39 BusyBody: creating ar
☑ 40 Disruption and recove
Step 2. Select your favorite articles
Articles in top list
Articles not in top list
ACES: A Cross-Discipline Platform and Method for Communication and Language Research
(M3) Articles of Coauthors
Groupware
by Carl Gutwin,
Rate the releva
Using Faceboo
by Moira Burke,
Rate the releva
Explore CSCW 2013 recommended articles!

Figure 17. Promotion in conference homepage in Conference Navigator, controllable interface.

The users that clicked on the link to use the Talk Recommendations had to complete 2 steps in order to obtain recommendations:

1) Select their favorite authors of papers from previous CSCW and CHI on a list of the 100 most cited authors of the CSCW and CHI conferences (as seen Figure 18), and,


CSCW
Home >> Recommendation of Talks
Home
Papers
People
My CSCW
Recommendations
Step 1: Choose Favorite Authors
Step 2: Choose Favorite Articles
Step 3: Explore Article Recommendations

Step 1 Select at least 3 authors relevant to your interests by clicking on the respective checkbox. These are the most cited authors -in the ACM DL- that have published in previous CSCW conferences.

Authors selected: Dan Cosley [x] Scott Counts [x] Shelly Farnham [x] Janet Go [x] Paul Resnick [x] Jeremiah Scholl [x] Junehwa Song [x]

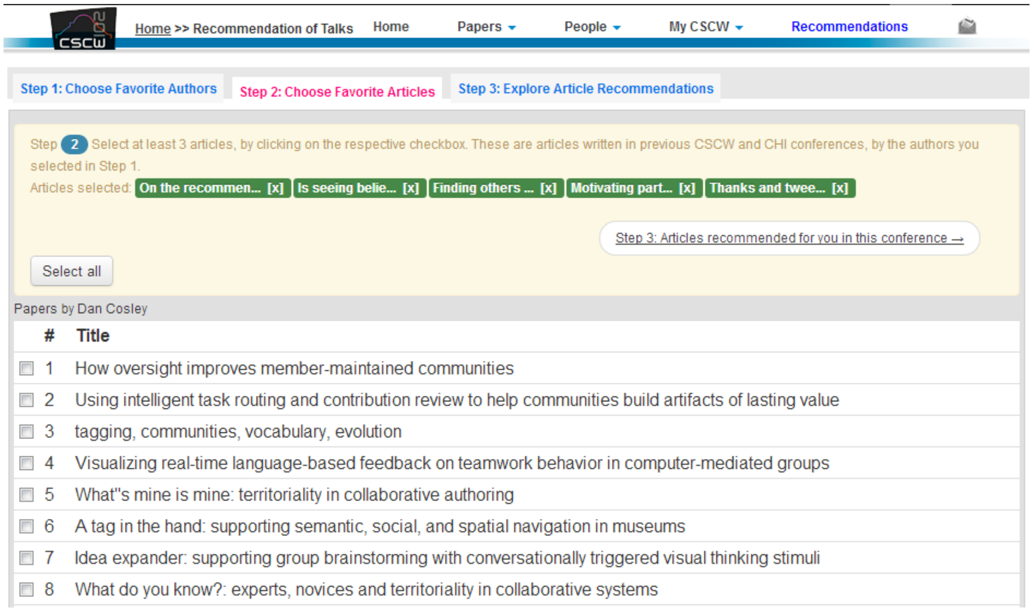
Step 2: Choose articles of selected authors →

Select all

	Author name
□	Joanna Abraham
□	Morgan Ames
□	Ofer Arazy
□	Elizabeth Bales
□	Richard Banks
□	Jeffrey Bardzell
□	Shaowen Bardzell
□	Abraham Bernstein
□	Jeremy Birnholtz

Figure 18. First step that CN3 users followed in the CSCW study: choose favorite authors

2) Choose the articles authored by these authors in CSCW and CHI conferences, which will be used to generate recommendations (as shown in Figure 19).


CSCW 2013
Home >> Recommendation of Talks
Home
Papers
People
My CSCW
Recommendations
Step 1: Choose Favorite Authors
Step 2: Choose Favorite Articles
Step 3: Explore Article Recommendations
Step 3: Articles recommended for you in this conference →
Select all

Step 2 Select at least 3 articles, by clicking on the respective checkbox. These are articles written in previous CSCW and CHI conferences, by the authors you selected in Step 1.

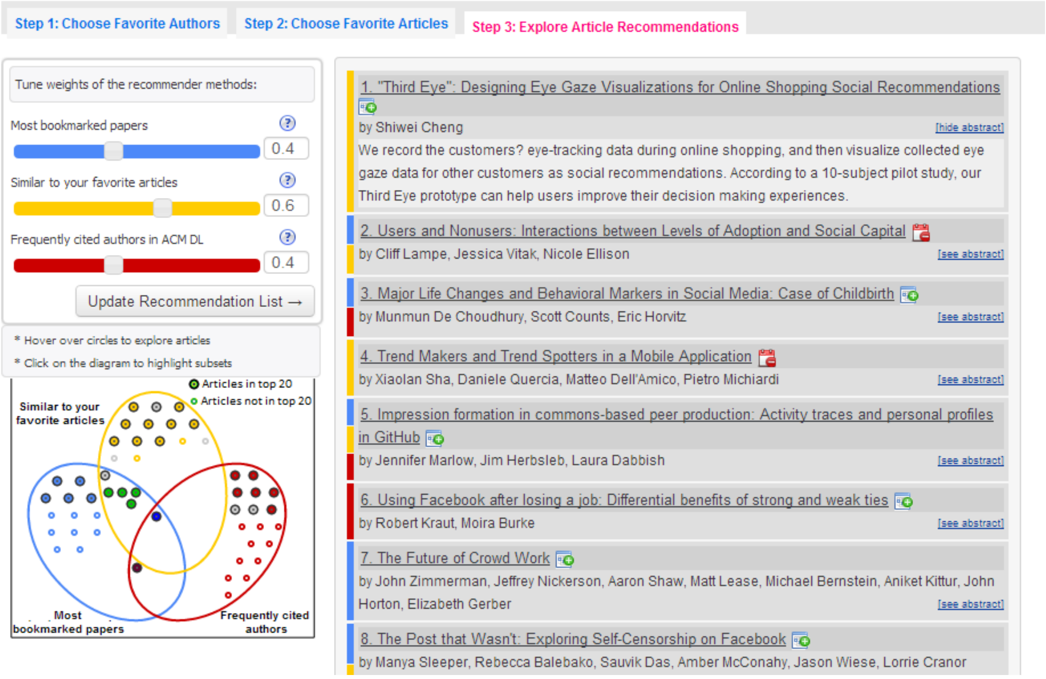
Articles selected: On the recommen... [x] Is seeing belie... [x] Finding others ... [x] Motivating part... [x] Thanks and twee... [x]

Papers by Dan Cosley

	#	Title
□	1	How oversight improves member-maintained communities
□	2	Using intelligent task routing and contribution review to help communities build artifacts of lasting value
□	3	tagging, communities, vocabulary, evolution
□	4	Visualizing real-time language-based feedback on teamwork behavior in computer-mediated groups
□	5	What''s mine is mine: territoriality in collaborative authoring
□	6	A tag in the hand: supporting semantic, social, and spatial navigation in museums
□	7	Idea expander: supporting group brainstorming with conversationally triggered visual thinking stimuli
□	8	What do you know?: experts, novices and territoriality in collaborative systems

Figure 19. Second step that CN3 users followed in the CSCW study: choose favorite articles

3) After the second step was completed, users saw the list of recommended talks in a controllable (Figure 20) or non-controllable recommender interface (Figure 21), depending on the condition they were assigned.


Step 1: Choose Favorite Authors | Step 2: Choose Favorite Articles | Step 3: Explore Article Recommendations

Tune weights of the recommender methods:

Most bookmarked papers ? 0.4

Similar to your favorite articles ? 0.6

Frequently cited authors in ACM DL ? 0.4

Update Recommendation List →

* Hover over circles to explore articles

* Click on the diagram to highlight subsets

1. "Third Eye": Designing Eye Gaze Visualizations for Online Shopping Social Recommendations

by Shiwei Cheng [hide abstract]

We record the customers? eye-tracking data during online shopping, and then visualize collected eye gaze data for other customers as social recommendations. According to a 10-subject pilot study, our Third Eye prototype can help users improve their decision making experiences.

2. Users and Nonusers: Interactions between Levels of Adoption and Social Capital

by Cliff Lampe, Jessica Vitak, Nicole Ellison [see abstract]

3. Major Life Changes and Behavioral Markers in Social Media: Case of Childbirth

by Munmun De Choudhury, Scott Counts, Eric Horvitz [see abstract]

4. Trend Makers and Trend Spotters in a Mobile Application

by Xiaolan Sha, Daniele Quercia, Matteo Dell'Amico, Pietro Michiardi [see abstract]

5. Impression formation in commons-based peer production: Activity traces and personal profiles in GitHub

by Jennifer Marlow, Jim Herbsleb, Laura Dabbish [see abstract]

6. Using Facebook after losing a job: Differential benefits of strong and weak ties

by Robert Kraut, Moira Burke [see abstract]

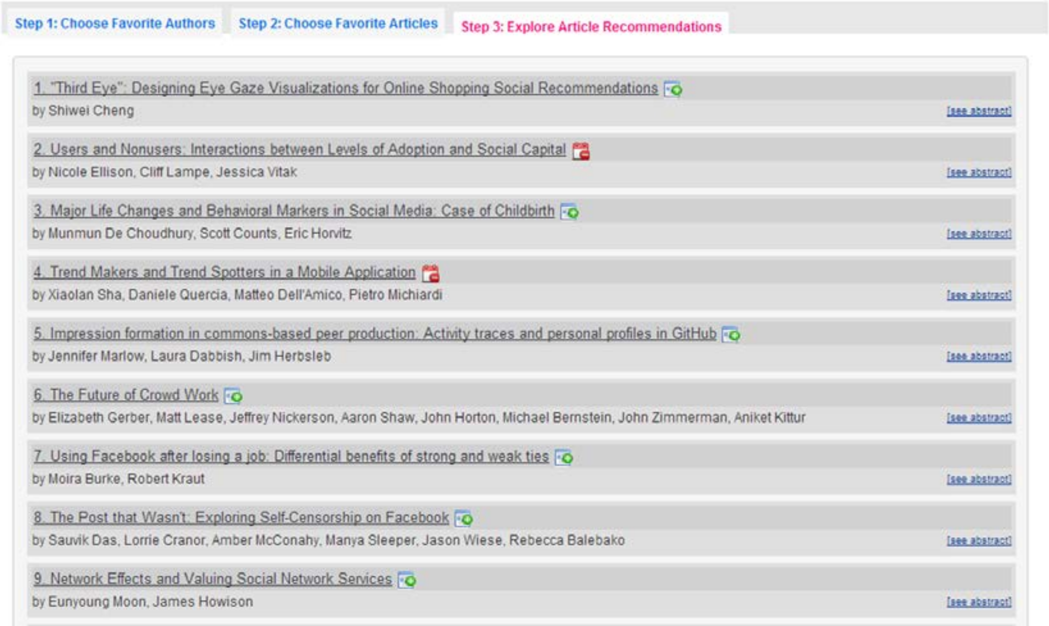
7. The Future of Crowd Work

by John Zimmerman, Jeffrey Nickerson, Aaron Shaw, Matt Lease, Michael Bernstein, Aniket Kittur, John Horton, Elizabeth Gerber [see abstract]

8. The Post that Wasn't: Exploring Self-Censorship on Facebook

by Manya Sleeper, Rebecca Balebako, Sauvik Das, Amber McConahy, Jason Wiese, Lorrie Cranor

Figure 20. Controllable recommender interface in the CSCW 2013 user study.


Step 1: Choose Favorite Authors | Step 2: Choose Favorite Articles | Step 3: Explore Article Recommendations

1. "Third Eye": Designing Eye Gaze Visualizations for Online Shopping Social Recommendations

by Shiwei Cheng [see abstract]

2. Users and Nonusers: Interactions between Levels of Adoption and Social Capital

by Nicole Ellison, Cliff Lampe, Jessica Vitak [see abstract]

3. Major Life Changes and Behavioral Markers in Social Media: Case of Childbirth

by Munmun De Choudhury, Scott Counts, Eric Horvitz [see abstract]

4. Trend Makers and Trend Spotters in a Mobile Application

by Xiaolan Sha, Daniele Quercia, Matteo Dell'Amico, Pietro Michiardi [see abstract]

5. Impression formation in commons-based peer production: Activity traces and personal profiles in GitHub

by Jennifer Marlow, Laura Dabbish, Jim Herbsleb [see abstract]

6. The Future of Crowd Work

by Elizabeth Gerber, Matt Lease, Jeffrey Nickerson, Aaron Shaw, John Horton, Michael Bernstein, John Zimmerman, Aniket Kittur [see abstract]

7. Using Facebook after losing a job: Differential benefits of strong and weak ties

by Moira Burke, Robert Kraut [see abstract]

8. The Post that Wasn't: Exploring Self-Censorship on Facebook

by Sauvik Das, Lorrie Cranor, Amber McConahy, Manya Sleeper, Jason Wiese, Rebecca Balebako [see abstract]

9. Network Effects and Valuing Social Network Services

by Eunyoung Moon, James Howison [see abstract]

Figure 21. Non-controllable recommender interface in the CSCW 2013 user study.

5.1.4 Exploratory Click-Log Analysis

Although the metrics described in the previous hypotheses combined provide support for differences between the non-controllable and the controllable interface if the expected sample sizes are met, it might be possible to find large differences among people in the same condition, for instance, numbers of clicks of different users in the controllable interface. The distribution of clicks on each component is analyzed to assess how differently people explore the control (sliders) and explainability (Venn diagram) widgets. By comparing the distribution of average number of clicks upon each component of the interface it might be possible to find patterns that describe the user behavior in the interfaces.

5.1.5 Post-conference survey

To avoid misinterpreting the user behavior represented by objective metrics such as *time spent in the task* as engaged behavior (when users might have a different perception about the system), further information was collected through a post-conference survey by asking CN3 users about the recommender interface. The survey questions were adapted from the user evaluation frameworks described in 4.2, and they are shown in Table 6.

Table 6. Questions of the survey to understand the user experience and engagement that will be conducted during CSCW 2013.

<table>
<tr><td colspan="2"><< 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>I felt in control of combining different recommendation methods by using the sliders.</td></tr>
<tr><td>*2</td><td>The ability to control the recommendation methods increased my satisfaction with the list of recommended talks.</td></tr>
<tr><td>*3</td><td>The ability to control the recommendation methods increases my trust in the list of recommended talks.</td></tr>
<tr><td>*4</td><td>When looking at the list of recommended talks I am interested to examine which recommendation method has been used.</td></tr>
<tr><td>*5</td><td>I think the Venn diagram visualization helped me to understand why a talk was recommended.</td></tr>
<tr><td>*6</td><td>I think the Venn diagram visualization was useful to identify talks recommended by a specific recommendation method or by a combination of recommendation methods.</td></tr>
<tr><td>*7</td><td>The ability to use the Venn diagram to examine the talks recommended increases my trust in the list of recommended talks.</td></tr>
<tr><td>8</td><td>I understood why the talks were recommended to me.</td></tr>
<tr><td>9</td><td>The items recommended matched my interests.</td></tr>
<tr><td>10</td><td>The items recommended were diverse.</td></tr>
<tr><td>11</td><td>I became familiar with the recommender interface very quickly</td></tr>
<tr><td>12</td><td>I lost track of time while I was using the recommender interface</td></tr>
<tr><td>13</td><td>Overall, I am satisfied with the recommender interface</td></tr>
<tr><td>14</td><td>The recommender made me more confident that I didn’t miss relevant talks</td></tr>
<tr><td>15</td><td>I would use this recommender system again for another conference in the future</td></tr>
<tr><td>16</td><td>I would suggest my colleagues to use this recommender system when they attend a conference in the future</td></tr>
</table>

17	I do not think that a social conference support system - like Conference Navigator- needs Talk Recommendation functionality
18	Comments and general feedback from the subject.

5.2 RESULTS

To assess the impact of our visual controllable interface we tracked a specific set of users' actions under both interfaces. In order to answer the research questions related to Study 1, we analyzed users' traces on the interfaces as well as the answers to the survey.

5.2.1 Usage and Engagement

In total, 168 people created their accounts and used Conference Navigator during the CSCW conference (this period spans from February 17th of 2013 to March 3rd of 2013). Out of the total number, 161 users were exposed to one or another recommendation condition (84 to visual and 77 to regular interface). Working with a recommender component was an option, not a requirement. The only engagement strategy used was the advertising of one or another recommender interface on the system home page. Out of 161 exposed users, 37 used the talk recommender. 22 people used the visual controllable interface and 15 the non-controllable interface. From this group of 37 people, 17 answered a survey to obtain feedback regarding their experience using the talk recommendation.

The participant's data provide some preliminary evidence that the visual interface was more attractive for the users. While the number of users exposed to both conditions was about the same, about 50% more users were sufficiently interested in the visual option to go through the three steps described (N=22) in section 5.1.3 until they could actually see the recommendations, compared to only 15 that used the interface with a static list of recommendations. Although the conversion rate of the controllable interface is larger (22 out of 84 = 26.2%) than in the non-controllable interface (15 out of 77 = 19.5%), this difference is not significant given a two-sample test for equality of proportions $\chi 2 = 0.678$, $p = 0.4103$. The effect size for the proportions of 26.2% and 19.5% is equal to 0.16; which is considered a small effect size given Cohen's guidelines (Prajapati, Dunne, & Armstrong, 2010). The required sample size to find a significant difference in a two-sample test for equality of proportions given $\alpha=0.05$, power of 0.8, and an effect size of 0.16 is equal to 613 subjects, a much larger value than the 37 users of this field study. The average numbers of rated talks (13 vs. 15) and bookmarked talks (4.67 vs. 4) per active users were comparable, as seen in Table 7.

Table 7. Usage metrics of each recommender interface: Condition A (visual controllable) and Condition B (static list)

Metric	Condition A	Condition B
# Users exposed to recommendations	84	77
# Users who used recommendations	22	15
# Users who answered survey	11	6
# Bookmarks / avg per user	28 / 4.67	32/ 4
# Users who bookmarked	6	8
# Papers rated / avg per user	130 / 13	45 / 15
# Users who rated	10	3
Average user rating	3.73	3.34

Similar evidence was provided by the analysis of user interaction with talks within the interfaces. As the data shows (Table 7), the visual interface engaged more users into the active work with the papers (vs. passive browsing) – 10 users (45.5% of 22) were rating talks in the visual interface vs. only 3 (20% of 15) in the regular one, although the difference in proportions was not significant given a two-sample test for equality of proportions $\chi^2 = 1.54$, $p = 0.21$. In this case, the effect size for two proportions of 45.5% and 20% is equal to 0.54. The required sample size to find a significant difference in a two-sample test for equality of proportions given α=0.05, power of 0.8 and an effect size of 0.54 is equal to 54 subjects, which is again, as in the previous test of proportions, larger than the 37 users of the recommender interfaces in this field user study.

5.2.2 Hypothesis Testing on User Engagement

Study 1 is conducted in a real setting, thus allowing the testing of user engagement between interfaces with usual action metrics (number of bookmarks, number of clicks) and also with others not available in a laboratory setting, such a number of people who return to see the recommender interface. For this reason the CSCW study was important to test user engagement and compare it between interfaces. Despite the small number of subjects, which did not provide enough power to test for all the hypotheses described in the research design–compared to what the power analysis suggested for some metrics–it was still possible to find some statistically significant differences. In the following subsections, hypotheses related to user engagement are presented.

Table 8. Engagement metrics of each recommender interface: Condition A (visual controllable) and Condition B (static list) (*p < 0.05 using an independent-samples Mann-Whitney U test)

Metric	Controllable	Non-controllable	p-value
Number of talks explored using the recommender interface (avg per user)	**16.84***	**7.27**	**0.045**
Number of talks explored excluding Venn diagram actions (avg per user)	**10.69**	**7.27**	0.577
Number of talks bookmarked using the recommender interface / avg per user	28 / 4.67	32/ 4	
Number of clicks in the recommender interface (avg per user)	**22.77***	**8.93**	**0.022**
Amount of time spent in the recommender interface (seconds)	261.72	192.2	0.098
Number of people who came back to recsys interface	7 (31.81%)	2 (13.33%)	

5.2.2.1 Hypothesis testing on number of talks explored

The variable "talks explored" considers actions in the interface directly related to inspecting talks in the conference: bookmarking, rating, opening the abstract, closing the abstract, hovering over Venn diagram circles (this last action is only available in the controllable interface). A larger number of talks is expected to be explored in the controllable interface, since this interface has explainable features (for each talk, individual colors and the Venn diagram representing recommendation methods) that should increase user attention on understanding why a talk is recommended. The following hypotheses were tested with respect to the number of talks explored:

$H_{1\text{-}0}$: There is no difference between user means on number of talks explored in the controllable and the non-controllable interface ($\mu_C = \mu_{NC}$)

In the controllable condition 19 out of 22 users explored talks using the controllable interface and 11 out of 15 people explored them with the non-controllable interface. A Mann-

Whitney U test instead of a t-test was used to test $H_{1\text{-}0}$ since the distributions violated normality, but both had similar shapes (positively skewed). We reject the null hypothesis, since a statistically significant difference was found in the number of talks explored between the controllable (M=16.84,S.E=3.01) and the non-controllable interface (M=7.27,S.E=2.06), p=0.045.

5.2.2.2 Hypothesis testing on number of talks explored excluding Venn diagram actions

The previous hypothesis might be considered biased towards the controllable interface since the actions on the Venn diagrams are only available in the controllable interface. Hence, this second hypothesis compares actions related to talk exploration available in both the controllable and the non-controllable interfaces: bookmarking, rate, opening the abstract, closing the abstract. Although with a smaller effect than the previous hypothesis, it would be expected that a larger number of talks would be explored in the controllable interface compared to the non-controllable one. The following hypotheses were tested with respect to the number of talks explored:

$H_{1.1\text{-}0}$: There is no difference between user means on number of talks explored (excluding actions on the Venn diagram) in the controllable and the non-controllable interfaces ($\mu_C = \mu_{NC)}$

A Mann-Whitney U test instead of a t-test was used to test $H_{1.1\text{-}0}$ since the distributions violated normality, but both had similar shapes (positively skewed). We failed to reject the null hypothesis, since no statistically significant difference was found in the number of talks explored (considering actions available in both interfaces) between the controllable (M=10.69,S.E.=2.36) and the non-controllable interfaces (M=7.27,S.E.=2.06), p=0.577.

5.2.2.3 Hypothesis testing on talks bookmarked

We compared whether there was a significant difference in the number of bookmarks that users made in the controllable versus the non-controllable conditions. The controllable interface would be expected to draw more user engagement in this metric, since the explainability features lead users to be more trusting, and then more confident about bookmarking talks. The null hypothesis is:

$H_{2\text{-}0}$: There is no difference between user mean number of talks bookmarked in the controllable and the non-controllable interface (μ_C = $\mu_{NC)}$

The null hypothesis was tested using an independent samples t-test. There was no significant difference in the mean number of talks bookmarked between the controllable (M=4.67,S.E.=1.74) and the non-controllable (M=4.0, S.E.=0.57) conditions, p = 0.690.

5.2.2.4 Hypothesis testing on clicks in the recommender interface

The variable “clicks on the recommender interface” considers all the possible actions over the interface. There is expected to be a significant difference in this variable between the two conditions since the controllable interface had 4 more types of actions available than the non-controllable interface, these being the visual widgets (change sliders, update recommendation list, hover over circle, click on ellipse), and we expect users to be motivated to explore the interface through these additional actions. The following hypotheses were tested with respect to the number of clicks on the recommender interface:

$H_{3\text{-}0}$: There is no difference between user means on the number of clicks in the controllable and the non-controllable interface (μ_C = $\mu_{NC)}$

A Mann-Whitney U test instead of a t-test was used to test $H_{3\text{-}0}$ since the distributions violated normality, but both had similar shape (positively skewed). We reject the null hypothesis because a statistically significant difference was found in the number of clicks between the controllable (M=22.77,S.E=3.73) and the non-controllable interface (M=8.93,S.E=1.82), p=0.022.

5.2.2.5 Hypothesis testing on time spent in the recommender interface

The time that users spent with each recommender interface was measured. Users had more actions to explore in the controllable interface and the explainable features should make them understand and meditate more over their decisions to choose specific talks; for that reason, users would be expected to spend more time in the controllable interface than the non-controllable one. The following hypothesis was tested with respect to this variable:

$H_{4\text{-}0}$: There is no difference between mean time spent on each interface in the controllable and the non-controllable interface ($\mu_C = \mu_{NC)}$

A Mann-Whitney U test instead of a t-test was used to test $H_{4\text{-}0}$ since the distributions violated normality, but both had similar shape (positively skewed). We failed to reject the null hypothesis, since no statistically significant difference was found in the time spent in the recommender interface between the controllable (M=261.73,S.E=56.65) and the non-controllable conditions (M=192.2,S.E=76.92), p=0.098.

5.2.2.6 Summary of Results on User Engagement

The results of the hypothesis testing on user engagement metrics provide preliminary evidence that the users of the controllable recommender were more attracted to the features of the controllable interface compared to users of the non-controllable condition: users significantly clicked more and explored more talks in the controllable interface. However, the result is not conclusive, since other important metrics that could have a stronger effect in the controllable interface, such as number of bookmarks, time spent on each interface, and the number of talks explored (considering actions available in both interfaces), were not significantly different between the two conditions. Moreover, a power analysis comparing proportions indicates that the small effect size in one metric (conversion rate of observing a promotion message and actually using the interface) and a small sample size in another variable (proportion of people who rated papers) hindered the chances of finding significant differences between interfaces.

5.2.3 Action Analysis

To understand the value of the interactive functionality to the system's users, we analyzed mouse clicks and hovers over different elements of the interface. A good overview of interactive feature usage is provided by Figure 22, which shows the average numbers of actions per user of each type of action.

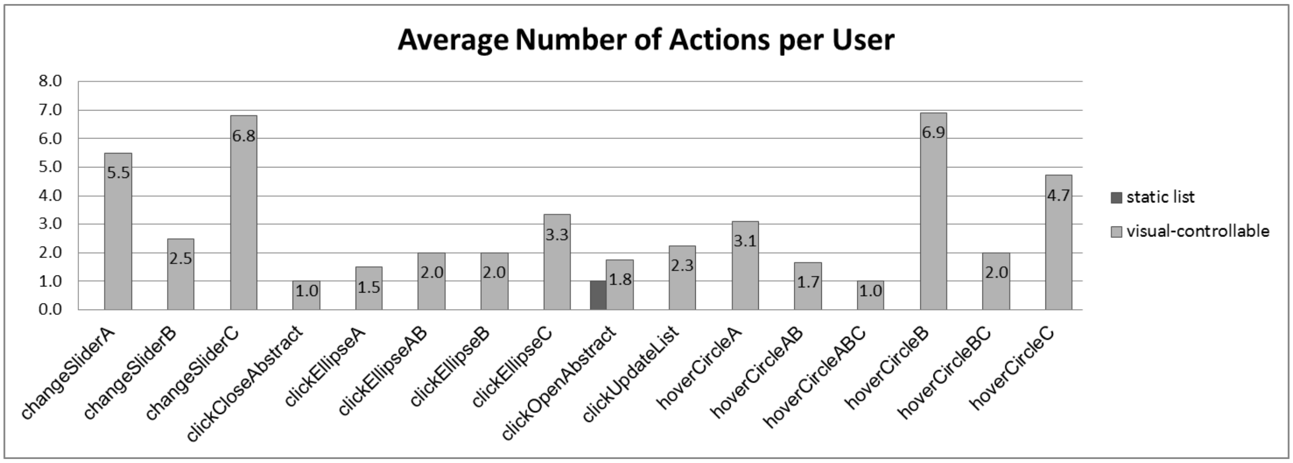


Figure 22. Average number of actions per user. The horizontal axis indicates the type of actions. The dark grey bars indicate actions in the static list interface, and the light grey actions in the controllable interface.

5.2.3.1 Sliders

On average, users clicked the button “click Update list,” which updates the recommended list, 2.3 times.

In terms of which method they preferred to use for moving the sliders, on average the users changed the slider of method C (articles of co-authors) 6 times, compared to 5.5 times for method A (papers of famous authors) and only 2.5 times for method B (content-based recommender). Further log analysis shows that for method A 22% of the times the weight was set to more than 0.5, compared to 80% of the time for slider B and 50% of the time for slider C.

5.2.3.2 Venn Diagram

Users could click over the large ellipses or hover over the circles representing papers of each method or intersection of methods. Users explored 6.9 papers on average in the B area of content-based papers, 4.7 papers from the C method “articles by co-authors” and only 3.1 papers from the A area of papers with high “author impact”. This is only slightly higher than the average number of papers explored in the intersection of methods B and C, which was 2, supporting the larger relevancy of methods B and C, and also providing more evidence of the Venn diagram’s role in identifying the papers recommended by both methods.

5.2.3.3 Impact

The usage data indicates that, in some aspects, the visual interface was more engaging to the users and that the users were actively using the offered interactive functionality. However, does it provide any advantage to its uses? We analyzed two impact factors – the average talk ratings and the average rank of relevant talks. A difference in average user rating might provide some evidence that one interface helps the users to find better talks. As the data shows, the average rating of ranked talks is, indeed, a bit higher in the visual controllable interface (mean average rating of 3.73) than in the non-controllable interface (mean average rating 3.34). A difference in the average rank of good talks is evidence that one system can better “push” good talks to the top of the ranked lists (lower ranks) where the talks have a better chance of being noticed. In our case, indeed, the average rank of papers rated 4 or 5 was 9.26 for the visual controllable condition and 11.78 for the static list of papers. Unfortunately we had too few users to run a statistical test with enough power to find significant differences, so both observations merely provide a hint, not reliable evidence in favor of the visual system.

5.2.3.4 Summary of the Action Analysis

The analysis of actions performed by users on the recommender interfaces allowed us to describe with more detail the user behavior, particularly on the controllable interface since it had a richer set of actions. From this analysis, we learned that users that tried the sliders and that passed their mouses over the Venn diagram tended to pay more attention to methods B (content-based) and C (articles of co-authors) compared to method A (articles by famous authors). Two metrics to compare the impact of the controllable interface over the non-controllable interface were tested (average ranking of scheduled papers and average rating), and they did not show a significant difference between both conditions.

5.2.4 Survey Analysis

After the conference had ended, an e-mail was sent to users that had used the recommender feature in Conference Navigator, inviting them to answer a survey about their experience using the talk recommendation feature. 11 people who had experienced the controllable condition answered the survey and 6 who had experienced the non-controllable condition. Table 9 shows the questions to which users had to answer how much they agreed with the statement presented.

Table 9. Results of the post-study survey (*=significant at p<0.05 using Mann-Whitney Test)

Question	Controllable recommender	Non-controllable recommender
1. The talks recommended matched my interests.	3.64±0.12	3±0.36
2. The talks recommended were diverse.	3.73±0.18	3.33±0.33
3. I became familiar with the recommender interface very quickly.	4.09±0.37	3.67±0.33
4. Following all the steps to generate a list of recommended talks required too much effort.	3.27±0.36	3.17±0.4
5. I lost track of time while I was using the recommender interface.	2.55±0.28	2.83±0.3
6. Overall, I am satisfied with the recommender interface.	3.18±0.20	3±0.36
7. The recommender made me more confident that I didn't miss relevant talks.	2.55±0.31	2.33±0.42
8. I would use this recommender system again for another conference in the future.	3.64±0.23	3.33±0.33

9. I would suggest my colleagues to use this recommender system when they attend a conference in the future.	3.36±0.24	3.33±0.33
10. I do not think that a social conference support system - like Conference Navigator- needs Talk Recommendation functionality	**2.36±0.23***	3.5±0.42
11. I understood why the talks were recommended to me.	3.45±0.24	2.83±0.47
Questions About the visual controllable recommender		
C.1. I felt in control of combining different recommendation methods by using the sliders.	3.64±0.16	-
C.2. The ability to control the recommendation methods increased my satisfaction with the list of recommended talks	3.36±0.31	-
C.3. The ability to control the recommendation methods increases my trust in the list of recommended talks	3.18±0.2	-
C.4. When looking at the list of recommended talks I am interested to examine which recommendation method has been used.	3.9±0.29	-
C.5. I think the Venn diagram visualization helped me to understand why a talk was recommended.	3.82±0.28	-
C.6. I think the Venn diagram visualization was useful to identify talks recommended by a specific recommendation method or by a combination of recommendation methods.	3.72±0.25	-
C.7 The ability to use the Venn diagram to examine the talks recommended increases my trust in the list of recommended talks.	3.18±0.33	-
Questions About the non-controllable recommender		
NC.1 I found it easy to adjust the list of talks recommended by changing my favorite authors and/or my favorite articles.		2.67±0.49

5.2.4.1 Summary of Results of the Survey

The survey indicates that people are more likely to think that a talk recommender system is necessary when using the visual controllable interface, which might be a result of being aware of why certain talks are recommended, i.e., the explainability of the controllable recommender interface. The high average agreement on question C.4 “When looking at the list of recommended talks I am interested to …” (3.9) supports the idea that the interface triggers a need for understanding how the talks were recommended, and the average rating of C.5 “I think the Venn diagram visualization helped me to understand why a talk was recommended" (3.82) suggests that the Venn diagram helps to support that need.

5.3 SUMMARY AND DISCUSSION

In this between-subjects study we contrasted the user engagement, behavior, and subjective experience of users who interacted with one of two possible recommender interfaces: controllable and non-controllable. We found initial but not conclusive evidence that the users were more engaged with the controllable interface than with the non-controllable one. Moreover, through a behavioral analysis, we gathered additional evidence of the usage of the controllable interface, finding that methods B (content-based recommendations) and C (papers by co-authors of favorite authors) attracted more user attention than method A (papers written by famous authors). Finally, the results of a post-study survey indicate that there is a small but significant effect on one area of users' perception about the interface: users are more likely to think that a recommender is necessary in Conference Navigator after using the controllable interface than after using the non-controllable interface.

5.3.1 Discussion

In terms of engagement, in this study we found some initial but not strong evidence that the users were more attracted to the controllable interface than to the non-controllable one. The percentage of users that clicked on the link at the home page to use the recommender was higher in the controllable interface, but the effect size was not large enough to find a significant difference with the non-controllable one. There was a significantly higher number of clicks using the controllable interface compared to the other one, but since the users of the controllable interface had a larger set of potential actions compared the non-controllable interface, this result might be biased.

The evaluation results also indicate that the visual controllable recommender had a positive effect on the user experience. Among the different user actions described in the behavioral analysis, users' hovering the mouse over the circles in the Venn diagram indicates that users had a real interest in exploring the recommendations, which is supported by the positive answers to the questions C.4, C.5, and C.6 in the survey (N=11).

One result that looks strange is that people were more likely to rate papers than bookmark them. We explain this behavior by the system's failure in leveraging people's feedback: we did not consider using the users' rating to re-generate recommendations. This system behavior could have negatively affected the users' trust in the system because their opinion was not used to update their recommendations.

Important lessons learned from this study led to some adjustments on the initial design of the second study:

(a) Separate rating and bookmarking tasks to avoid users' confusion: log analysis revealed that 3 users that entered the visual recommender interface rated the whole list of 20 papers presented on the screen. Then, they clicked on the "*Update Recommendation List*" button, and after that action they simply left the visual recommender page. The explanation for this behavior is that the recommender algorithm was not considering the ratings as part of the user feedback to update the recommendations. The user pattern suggests that users lost interest in the recommender for not considering their preferences. Although this seems to be minor evidence, a similar behavior is described in the experiment conducted by Hijikata et al. (2012) that showed a decreased users' satisfaction when they were allowed to intervene a system but their feedback was not considered to update the list of recommendations. This motivated a change in the design of Study 2, where ratings were removed from the interface to allow users only the bookmarking of talks, and the action of rating was re-assigned to a different task and interface.

(b) Improve the filtering when users click on Venn Diagram: The interaction design of the Venn diagram, as explained in 3.2.1.3, allowed users to filter the papers on the item list by hiding all the papers except those whose area of the Venn diagram the user had clicked on, e.g., if users clicked on the content-based ellipse, only those papers would remain visible in the recommended list. However, the positions of the papers were kept fixed and a paper ranked 30 would require page scrolling to be visible. After considering user feedback, the list not only hid the non-selected papers, but also collapsed the list of selected papers, allowing the users to avoid unnecessary scrolling. This filtering interaction design was very appreciated by the users when they were performing the tasks of Study 2.

5.3.2 Implications

Despite the inconclusive results, there are three aspects that support the role of Study 1 in underpinning the implementation of Study 2. The results of this study allowed tuning the hybrid recommender parameters for Study 2: Which combination of weights in the hybrid recommender and which parameters in the content-based recommender provide the most accurate recommendation list? This question was answered based on the user preference feedback obtained from Study 1. Second, Study 1 log analysis supported the decision to separate the actions of bookmarking and rating into two separate tasks in the Study 2. Finally, the log analysis and user feedback helped to add a new feature for filtering after the users clicked in the intersection areas of the Venn diagram.

6.0 STUDY 2

A within subjects controlled laboratory study was conducted from April 26th through June 4th of 2013. Its purpose was to address every research question, but in particular research questions RQ3 and RQ4, related to the effect of user characteristics upon user engagement and user experience with the recommender interfaces, since this information was not possible to collect in Study 1.

6.1 DESIGN

6.1.1 Participants

Subjects were recruited by e-mail and by printed ads posted at the School of Information Sciences and the School of Computer Science at 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 at the University of Pittsburgh, as well as to students of 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 degree-in areas related to the iConference (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 avoid additional effects on their judgments of how relevant papers were, 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 to obtain the users' article preferences to generate recommendations for the other two conferences (following the aforementioned example, recommendations would be generated for iConference 2011 and 2013).

6.1.2 Hypotheses for Study 2

In this section I describe the hypotheses addressed in this study. In the case of evaluating user engagement, given the nature of the controlled laboratory setting, the hypotheses related to studying post-session behavior (number of times that the user comes back to the interface) are not considered in this study.

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

H1. Having more control will result in users more engaged, with the following objective metrics being larger than in a non-controllable interface:

1. Number of talks explored using the recommender interface
2. Amount of time spent in the recommender interface (seconds)

Only these two metrics are considered since they are comparable between both interfaces.

H2. Having a more controllable interface will result in users more engaged, with subjective measures being larger than in a non-controllable interface. This will be measured with survey questions to evaluate the following dimensions:

1. Focused attention,
2. Perceived usability,
3. Endurability, and
4. Novelty

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

H2. A user provided with more control will have a better user experience than a user with less control in a recommender system, as measured in higher average rating, higher precision@k, higher MAP, and higher nDCG.

H3. A user provided with more control will have a better user experience in a recommender system, as measured in perceived systems qualities, use beliefs, attitudes, and intentions.

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

H4. User characteristics such as Domain Knowledge, Familiarity with the Conference, Familiarity with the System, Trusting Propensity, and User Experience with Recommendation Systems are expected to interact with controllability to affect the user engagement with a recommender system

- User expertise in her own domain: Users with better domain knowledge will be more engaged with the more controllable interface, since their expertise will allow them take better advantage of the controls and inspectable visualization.
- Off-line community status: Users less engaged with the iSchools community might know fewer authors in the community, and as this is an important source for producing recommendations, it will decrease their likelihood to be engaged with the recommender system.
- Familiarity with the system: Users familiar with the system are more likely to interact with it. However, it isn’t clear which specific widgets they interact more frequently with, since they might prefer to use known features rather than new visual controls.
- Trusting propensity: Users with more trusting propensity will show smaller differences in engagement between the two interfaces.
- User experience with recommenders: People with user experience in recommender will be more curious about the results for controlling and inspecting recommendations methods, which will make them more likely to engage with the system.

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

H5. User characteristics such as User Expertise in her Domain, Familiarity with the iConference, Familiarity with the System, Trusting Propensity and User Experience with Recommendation Systems are expected to interact with controllability to affect the user experience with a recommender system

- User expertise in her own domain: Users with better expertise in their domains will have a better user experience with the controllable interface, since their expertise will allow them to take better advantage of the controls and inspectable visualization.
- Offline Community Status: Users more familiar with the iSchools community know more authors in the community, and as this is an important source to produce recommendations, it will improve the user experience with the system.
- Familiarity with the system: Users that have not used the system must become familiar with all the elements of the interface, but those who have used Conference Navigator before need only learn the new visual tools. This can have either a positive effect, if they consider that these tools enhance the interface for the purpose of bookmarking, or it can make users to ignore the visual tools to avoid cognitive strain in order to finish the task of bookmarking relevant papers, as suggested by Albers (1997).
- Trusting Propensity: Users with more trusting propensity will show smaller differences in user experience between the two interfaces, unlike users with small trusting propensity, who will prefer the more controllable interface.
- User Experience with Recommenders: People with experience with recommenders will be more curious about the results of controlling and inspecting recommendations methods, which will make them more likely to have a better user experience with the system.

6.1.3 Study 2 Procedure

The workflow of the study is presented in Figure 23. The diagram shows a user who starts by answering a pre-survey and then the two arrows indicate that she is assigned to one of two possible orders to begin the “Bookmarking” task: she can start with the non-controllable (No-C) interface and then continue with the controllable (C) interface or vice versa.

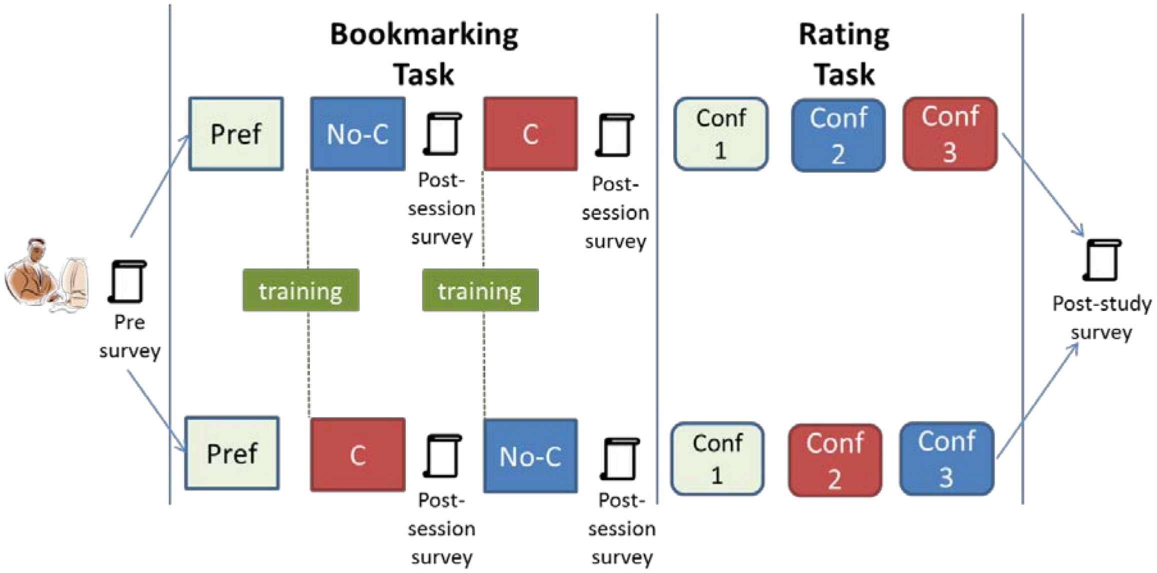


Figure 23. Workflow of Study 2. After answering the entry questionnaire, the subject is assigned to one of 2 possible sequences: No-controllable and then controllable interface, or vice versa.

Before any of the two recommender interfaces (No-C or C) is shown to the user, we obtain the initial user preference (Pref) by asking her to examine all the papers from the proceedings of a conference and bookmark the ones that she finds relevant, without a minimum or maximum limit on the number of papers to be bookmarked. Then she proceeds with the recommender interfaces (each one recommends papers from different conferences), but before she moves to the next recommender interface, she must answer a post-session survey. Finally, the user is given the “Rating” task, where she will rate (in a scale from 1 to 5) the relevance of all the papers that she saw in the “bookmarking” task. At the very end, the user has to answer a post-study survey with the purpose of comparing both recommender interfaces.

The subjects were assigned to the controllable and non-controllable conditions in such a way as to ensure an appropriate balance among the conditions in terms of the conference used for seeding (to obtain user's preference and produce content-based recommendations). Likewise, the subjects were assigned to the different conditions so as to balance the order (sequence) in which controllable and non-controllable interfaces were presented. Table 10 presents a summary of the number of subjects under each condition:

Table 10. Distribution of the 40 valid participants over the different conditions.

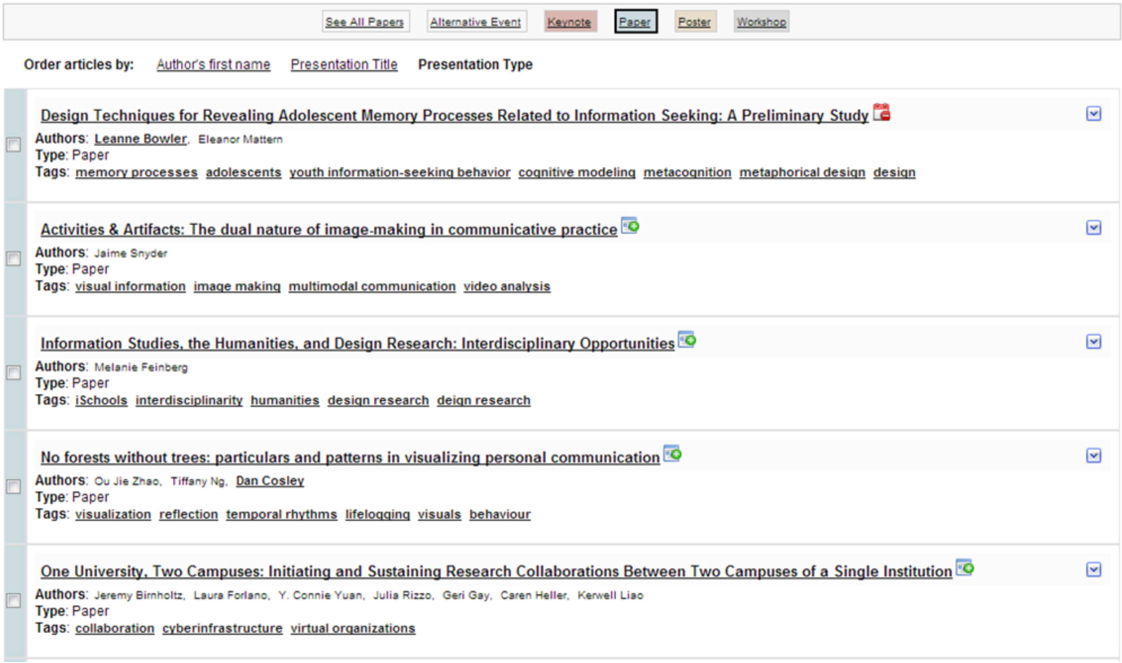
Seed Conference	Controllable Conference	Non-Controllable Conference	Sequence Order	
			Controllable => Non-controllable	Non- Controllable => Controllable
2011	2012	2013	3	3
2011	2013	2012	3	3
2012	2011	2013	4	4
2012	2013	2011	3	3
2013	2011	2012	4	4
2013	2012	2011	3	3

A more detailed description of the study workflow follows:

(a) Participants signed an informed consent of the benefits and risks of the study, as specified in the IRB.

(b) Participants filled 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 is given a hypothetical situation where she is attending the iConference and one week before the conference takes place her advisor ask her to identify the most relevant papers for her as well as for colleagues in the same laboratory. Participants are told that the main purpose of including “colleagues” is to increase the number of relevant papers, because the iConference is very diverse and there is a chance that too few papers are relevant only for her.



□ Design Techniques for Revealing Adolescent Memory Processes Related to Information Seeking: A Preliminary Study
Authors: Leanne Bowler, Eleanor Mattern
Type: Paper
Tags: memory processes adolescents youth information-seeking behavior cognitive modeling metacognition metaphorical design design

□ Activities & Artifacts: The dual nature of image-making in communicative practice
Authors: Jaime Snyder
Type: Paper
Tags: visual information image making multimodal communication video analysis

□ Information Studies, the Humanities, and Design Research: Interdisciplinary Opportunities
Authors: Melanie Feinberg
Type: Paper
Tags: iSchools interdisciplinarity humanities design research deign research

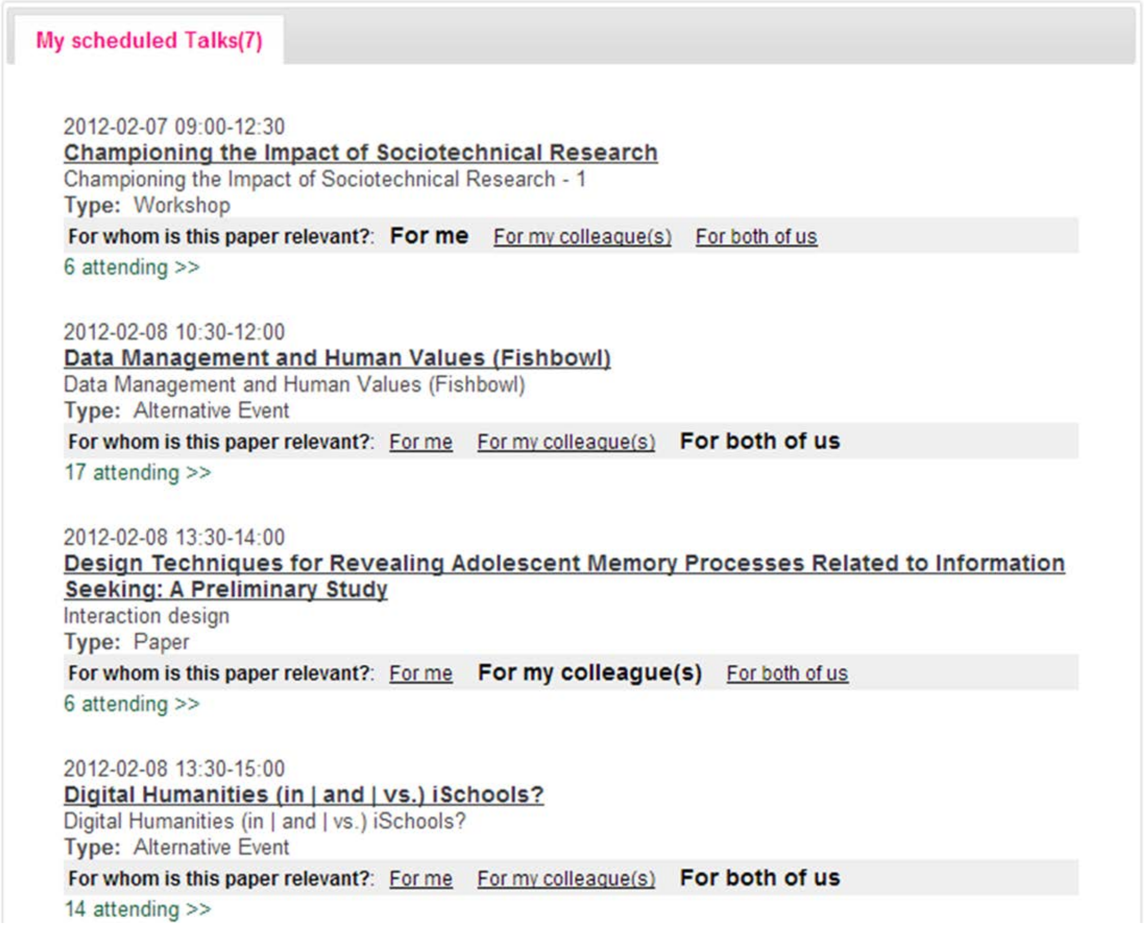
□ No forests without trees: particulars and patterns in visualizing personal communication
Authors: Ou Jie Zhao, Tiffany Ng, Dan Cosley
Type: Paper
Tags: visualization reflection temporal rhythms lifelogging visuals behaviour

□ One University, Two Campuses: Initiating and Sustaining Research Collaborations Between Two Campuses of a Single Institution
Authors: Jeremy Birnholtz, Laura Forlano, Y. Connie Yuan, Julia Rizzo, Geri Gay, Caren Heller, Kerwell Liao
Type: Paper
Tags: collaboration cyberinfrastructure virtual organizations

Figure 24. Proceedings page, where subjects perform the first subtask in the iConference study.

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 (Figure 24). After this step is finished, in a different screen that only shows the papers bookmarked, the user must state for each paper if it was judged as relevant only for her, for her and her colleagues, or only for her colleagues (Figure 25).



My scheduled Talks(7)

2012-02-07 09:00-12:30
Championing the Impact of Sociotechnical Research
Championing the Impact of Sociotechnical Research - 1
Type: Workshop
For whom is this paper relevant?: **For me** For my colleague(s) For both of us
6 attending >>

2012-02-08 10:30-12:00
Data Management and Human Values (Fishbowl)
Data Management and Human Values (Fishbowl)
Type: Alternative Event
For whom is this paper relevant?: For me For my colleague(s) **For both of us**
17 attending >>

2012-02-08 13:30-14:00
Design Techniques for Revealing Adolescent Memory Processes Related to Information Seeking: A Preliminary Study
Interaction design
Type: Paper
For whom is this paper relevant?: For me **For my colleague(s)** For both of us
6 attending >>

2012-02-08 13:30-15:00
Digital Humanities (in | and | vs.) iSchools?
Digital Humanities (in | and | vs.) iSchools?
Type: Alternative Event
For whom is this paper relevant?: For me For my colleague(s) **For both of us**
14 attending >>

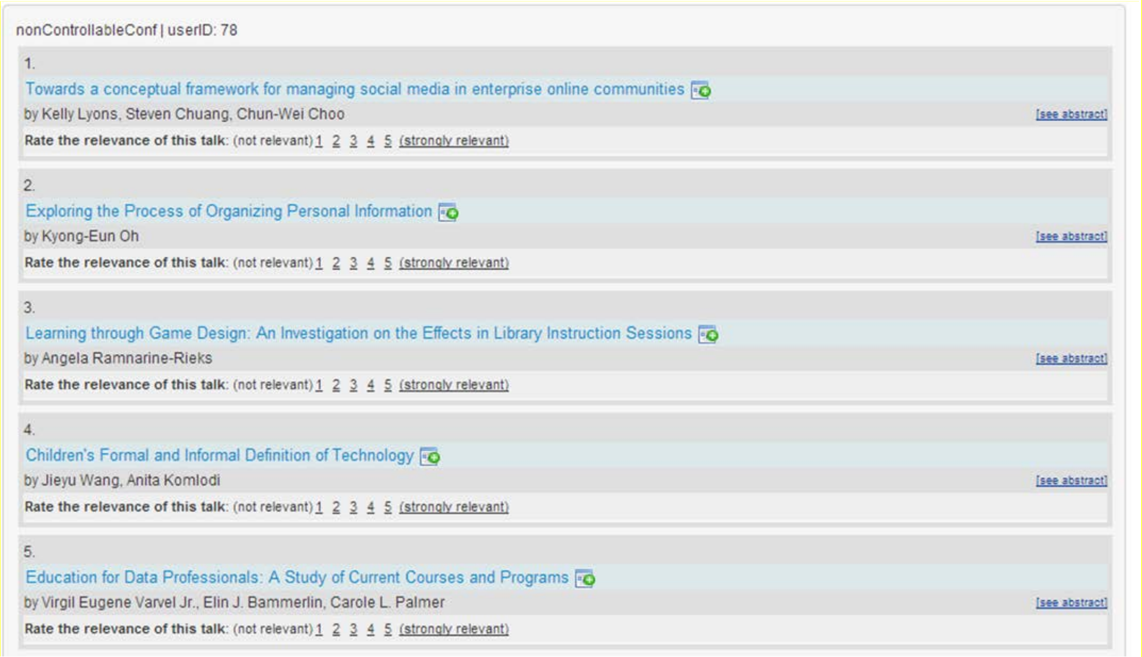
Figure 25. Screenshot shows bookmarked talks where the user must judge for whom they are relevant: a) for me, b) for my colleagues(s), c) for both of us.

2. Subtask 2: Using the controllable or the non-controllable recommender interface – depending on the condition the user was assigned to- 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,

using the same interface shown in Figure 25. After finishing this step, the use must answer a post-session survey regarding the interface (controllable or not) she was assigned in the first task.

3. Subtask 3: The subtask 2 is repeated for the other recommender interface – controllable or not-controllable- depending on the condition the user had been assigned to. At the end of this subtask, the use must answer a post-session survey regarding the interface (controllable or not) she was assigned in the second task.

(d) Task 2: Rate papers on a scale from 1 to 5, where 1 means not relevant at all and 5 means strongly relevant (Figure 26). The papers are sorted randomly and the icon besides the title indicates if the paper was bookmarked in the previous task. The subtasks are as follows:



nonControllableConf | userID: 78

1.
Towards a conceptual framework for managing social media in enterprise online communities
by Kelly Lyons, Steven Chuang, Chun-Wei Choo [see abstract]
Rate the relevance of this talk: (not relevant) 1 2 3 4 5 (strongly relevant)

2.
Exploring the Process of Organizing Personal Information
by Kyong-Eun Oh [see abstract]
Rate the relevance of this talk: (not relevant) 1 2 3 4 5 (strongly relevant)

3.
Learning through Game Design: An Investigation on the Effects in Library Instruction Sessions
by Angela Ramnarine-Rieks [see abstract]
Rate the relevance of this talk: (not relevant) 1 2 3 4 5 (strongly relevant)

4.
Children's Formal and Informal Definition of Technology
by Jieyu Wang, Anita Komlodi [see abstract]
Rate the relevance of this talk: (not relevant) 1 2 3 4 5 (strongly relevant)

5.
Education for Data Professionals: A Study of Current Courses and Programs
by Virgil Eugene Varvel Jr., Elin J. Bammerlin, Carole L. Palmer [see abstract]
Rate the relevance of this talk: (not relevant) 1 2 3 4 5 (strongly relevant)

Figure 26. Screenshot of the rating interface.

1. Subtask 1: The user must rate all the papers shown in the proceedings page of the seeding conference.
2. Subtask 2: The user must rate the set of all the recommended papers (shown in several lists in the controllable interface and in only one list in the case of the non-controllable interface) of the second subtask of task 1 “bookmark relevant papers.”
3. Subtask 3: The user must rate the set of all the recommended papers of the second subtask of task 2 “bookmark relevant papers.”

(e) Post-questionnaire: Obtain participants’ perception 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 finished 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 one.

6.1.4 Surveys in Study 2

Given the current literature on user-centric recommendations systems, the pre-survey captures user characteristics, beliefs and expectations. Study 2 (iConference study) required a pre-survey to obtain information about the subject’s personal characteristics and experience. ‘Trusting Propensity’ has been adapted from (Knijnenburg, Rao, & Kobsa, 2012). The questions on this survey can be examined in Appendix A.1.

The user answered two post-session surveys, one under each interface. The questions in these surveys can be examined in Appendix A.2 (for the non-controllable interface) and Appendix 0 (for the controllable interface). There are 10 questions in common, but also 7 questions that refer only to features in the controllable interface. Finally the post-study survey captures the subjective opinions of the users on comparing both interfaces. The questions can be read in Appendix 0.

6.2 RESULTS

This section is composed of three parts. First, it describes demographics and characteristics of the participants of the study. Then, it reports on the factor analysis conducted to obtain user characteristics based on the pre-study questionnaire. The third and final part addresses the research questions through regression analysis.

Table 11. General descriptive statistics of usage in the iConference study

	Non-controllable recommender	Controllable recommender
# Subjects	40	40
# Total bookmarks	638	625
# Average bookmarks/user	15.95	15.63
# Average rating	2.48±0.089	2.46±0.076

6.2.1 Participants' Descriptive Statistics

42 people participated in the study, but only 40 were considered valid subjects for the remaining analysis. One of the subjects was removed from the list of valid participants because he left the study before finishing the first task. Although being a PhD student in Information Science, he stated that he did not find any paper of interest in the proceedings of iConference. The other subject was removed from the list of participants because he was observed not paying enough attention to the articles' information (title, author lists', keywords, or abstract) in order to judge their relevance. In summary, he spent only 45 minutes to finish the study, clearly below the minimum time needed to finish the study, which on average took 90 minutes per user.

36 out of the 40 subjects are PhD students with the following breakdown by area: Information Science (16), Library Science (9), Computer Science (6), Telecommunications (3), Public Policy (1), and Public and International Affairs (1). The other 4 subjects were a Postdoctoral researcher in Computer Science, a research scientist with a PhD in Information Science, an assistant professor at a local college in Pittsburgh in Computer Science, and a young research intern in the PAWS lab with a BSc in Telecommunications. In terms of gender distribution, 17 were female and 23 were male. The age of the participants ranged from 21 to 62 years old, with a mean of 31.75 and a standard deviation of 6.5 years. 25% of the subjects are native speakers (10). Among the PhD students, 4 have just started recently, 11 have finished the preliminary examination, 12 have finished the comprehensive exam or qualifiers, and 7 have already defended their proposal, and 2 are about to give their final defense. In terms of publications, 7 subjects still have not published anything, the rest at least have published 1 or 2 conference or journal papers. 16 subjects have already served as reviewers for a workshop, a conference or a journal paper.

Most of the participants had not attended any iConference, almost 75% (N=29). In terms of engagement with the iSchools community, 14 people report agreeing or strongly agreeing with the statement "I feel engaged with the iSchools community", whereas 7 people disagree or strongly disagree and the majority, 19 subjects, feels neutral about that statement. A 30% of the subjects (N=12) had used Conference Navigator before the study.

In the three statements referring to trusting propensity, the answers were clearly skewed to agreeing or strongly agreeing with the statements. With the statement "*In general, people really do care about the well-being of others*" 33 people agrees or strongly agrees, with "*The typical person is sincerely concerned about the problems of others*" 26 people shows the same preference, and with "*Most of the time, people care enough to try to be helpful, rather than just looking out for themselves*" 32 people shows the same agreement. Only the second of the 3 aforementioned questions shows a less skewed distribution, with 12 subjects feeling neutral about the statement and 2 of them disagreeing with it.

Finally, in the three statements referring to familiarity with recommender systems, most subjects showed a high level of familiarity. 33 users agree or strongly agree with the statement "*I am familiar with online recommender systems*", 34 subjects expressed agreement or strongly agreement with the statement "*I have occasionally followed the advice of a recommender system*", but a less skewed distribution is observed for the statement " *I know of one or more methods used to produce recommendations in a system*", where 28 subjects agree or strongly agree, 9 expressed neutrality and 3 disagreed with the statement.

6.2.2 Factor Analysis: User Characteristics from Pre-Study Survey

The questions of the pre-survey are listed on the Appendix A.1. Subjects answered 19 questions, the first 4 about demographics (occupation/program, gender, native English speaker, age), with the other 15 intended to asses 5 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 (Knijnenburg, Bostandjiev, et al., 2012; Knijnenburg et al., 2011; Pu et al., 2011).

Table 12. Questions and their loadings on the latent factors resultant of the EFA. Maximum likelihood estimation used as factor extraction method and varimax rotation.

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

The exploratory factor analysis showed that some questions did not load well on any of the models fitted, so they were removed. For instance, the question related to the number of workshop papers published did not load well with the questions assessing the participant's research expertise. It loaded always partially onto different latent factors, so it was removed. The factor analysis model had 4 factors, and their respective loadings can be seen in Table 12. 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.

The variables considered in this study to control for the effects of user characteristics are:

- Occupation (PhD student in different areas, Postdoc, researcher, lecturer)
- Age (continuous variable)
- Gender (male/female)
- Research Expertise: construct of two standardized variables
- Trusting Propensity: construct of three standardized variables
- Experience with Recommender Systems: construct of two standardized variables
- Trust in Recommender Systems: standardized variable of one question
- Previous use of Conference Navigator (yes/no)

6.2.3 Statistical Analyses

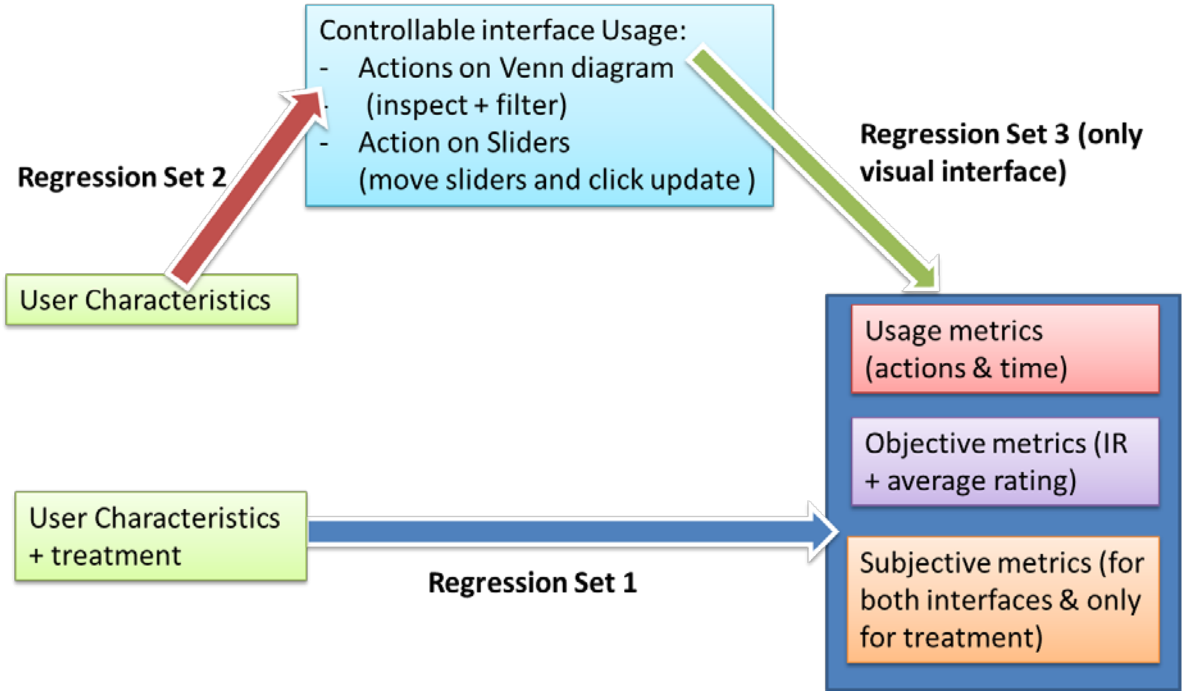

Controllable interface Usage:
- Actions on Venn diagram (inspect + filter)
- Action on Sliders (move sliders and click update)
Regression Set 2
Regression Set 3 (only visual interface)
User Characteristics
Usage metrics (actions & time)
Objective metrics (IR + average rating)
User Characteristics + treatment
Regression Set 1
Subjective metrics (for both interfaces & only for treatment)

Figure 27. Block model summarizes the statistical analysis conducted in Study 2.

The statistical analysis conducted to test the hypotheses considers three sets of regressions, as seen in Figure 27. The first set (regression set 1), conducts regressions on three groups of metrics (usage metrics, objective metrics, and subjective metrics) controlling for the treatment (the recommender interface) and for user characteristics. The other two sets of regressions focus only on the controllable interface to explain which features particular to the treatment influenced the experience of the user. The second set of regressions (regression set 2) analyses the actions with the Venn diagram and the actions using the sliders controlling for user characteristics. The final set (regression set 3), analyses the variables of the three dimensions of metrics, controlling for the number of actions with the Venn diagram and the sliders.

6.2.3.1 Regression Set 1: Regressions Controlling for User Characteristics, Interface, and Order

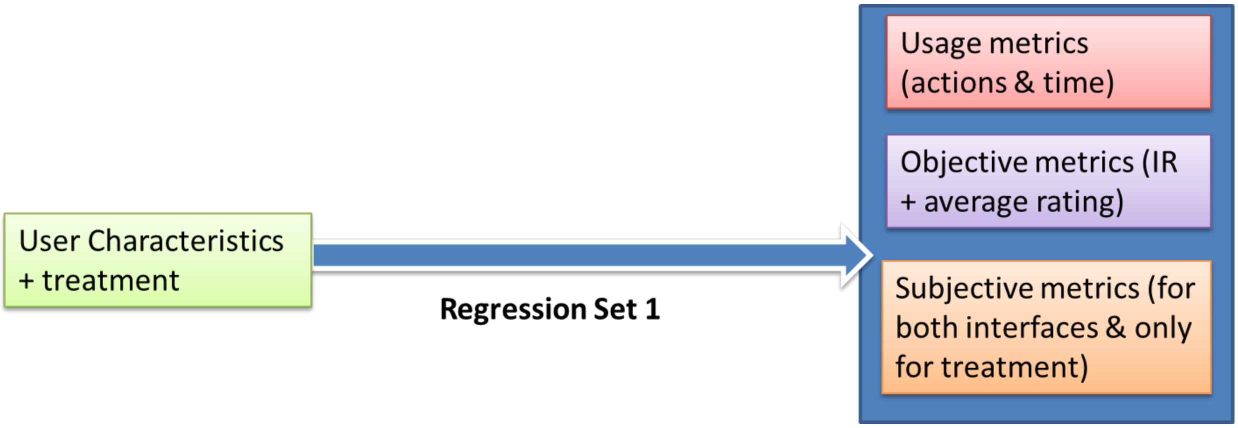

Usage metrics (actions & time)
Objective metrics (IR + average rating)
User Characteristics + treatment
Regression Set 1
Subjective metrics (for both interfaces & only for treatment)

Figure 28. Block model summarizes first set of regressions.

Regressions on Usage metrics

The two usage metrics considered, shown in Table 13, in this analysis are comparable between conditions: (a) Comparable talks explored (not considering Venn diagram hover actions), and (b) Time Spent. 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.

Table 13. Significant effects of the regressions on usage metrics

Metric	**Significant effects [(+) positive / (-) negative]**
Talks explored (comparable)	(-) experience with recsys, B = -0.1, p = 0.033 / exp(-0.1) = 0.91 (+) experience with CN, B = 0.31, p = 0.044 / exp(0.31) = 1.36
Time Spent	(-) order, B = -0.54, p < 0.001 / exp(-0.54) = 0.58 (+) condition*order, B = 0.6, p = 0.002 / exp(0.6) = 1.82

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 explore 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 on 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 increases the number of actions 36% compared to those not familiar, p = 0.044. Given that the list of items is presented in a similar design to other pages in the system (proceedings, top items, etc.), users familiar with the system will be more likely to try the analog features to explore talks since they know what to expect from them, decreasing their cognitive strain in exploring the items (Albers, 1997).

The treatment (controllable interface) did not have a significant effect on the number of explored talks, but it had a significant interaction with order, i.e., users that are shown the controllable interface after the non-controllable one spend significantly more time with the recommender. More specifically, 82% more time compared to the non-controllable interface when it is shown in the first place. Figure 29 shows this effect, where the time spent with the controllable interface (green line) is almost invariable whether presented first or second, yet the non-controllable interface (blue line) shows a significant drop in time spent when presented after the controllable interface.

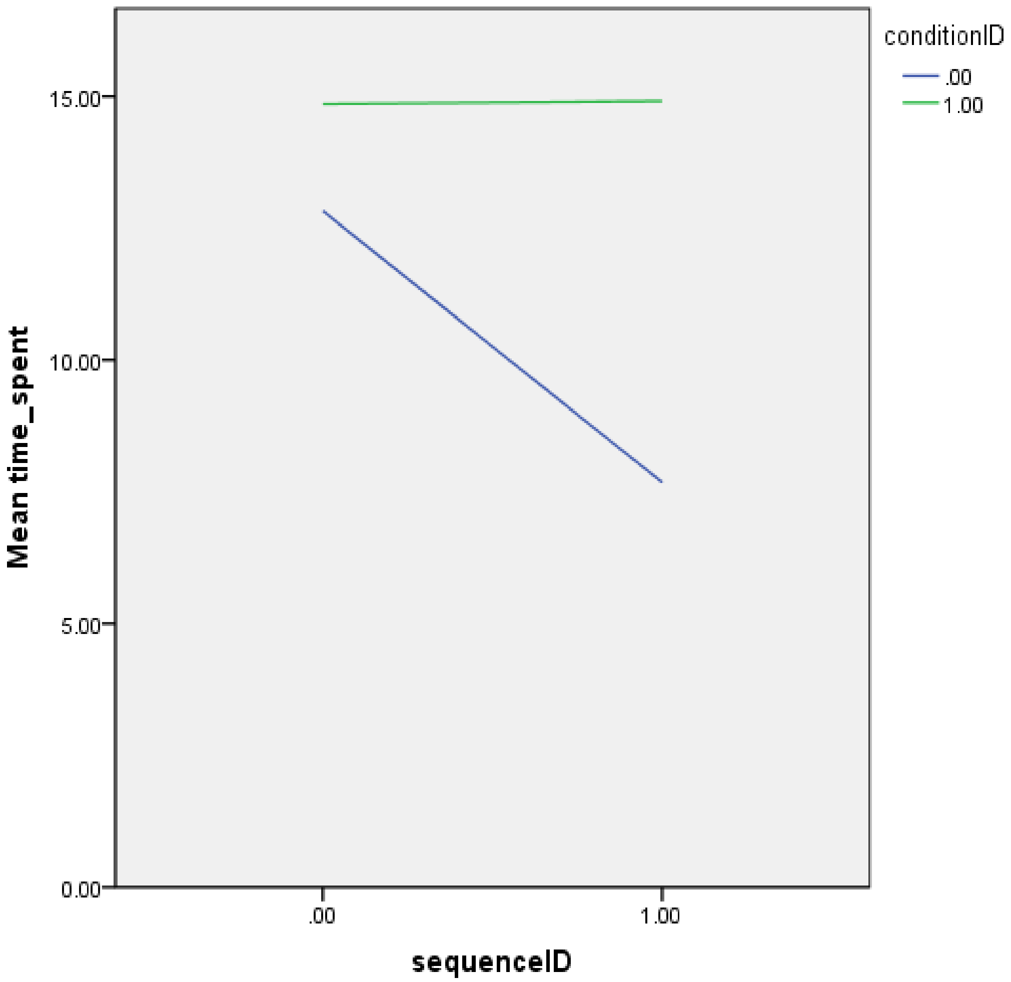


Figure 29. Effect of the treatment interface (conditionID = 1) and the order in which it was presented (sequenceID, x-axis) on the amount of time (y-axis) in minutes, that users spent inthe bookmarking task.

This can be interpreted either as a good or bad sign of engagement (users can be engaged with the features but also feel confused with the interface), therefore, subjective metrics are later used to understand this result.

Regressions on IR metrics

Table 14. Significant effects of the regressions on accuracy and IR metrics

Metric	**Significant effects [(+): positive / (-): negative]**
Average User Rating	No significant covariates
MAP	(+) conditionID, B = 0.08, p = 0.016 (+) BS in TELECOM, B = 0.24, p = 0.025
MRR	(+) order, B = 0.21, p = 0.021
nDCG	(+) programBS in Telcom, B = 0.18, p = 0.029
Precision	(+) programGSPIA PhD St, B = 0.18, p = 0.001
Precision@3	(+) programBS in Telcom, B = 0.54, p = 0.01
Precision@5	(+) programBS in Telcom, B = 0.45, p = 0.005 (+) gender(male), B = 0.12, p = 0.019

Linear mix-model regressions were conducted to understand the effect of the treatment (recommender interface) and user characteristics on average user rating, MRR, MAP , nDCG,, precision, and precision@n (with n=3 and n=5). The significant effects found on these regressions are shown in Table 14.

Neither the interface nor the user characteristics had a significant effect on explaining the variability of the average user rating, but the controllable interface produced an increase of 0.08 in the MAP compared to the non-controllable one. In addition, there was one user in the study with low expertise in her domain, currently in a master program of Telecommunications. Keeping all the variables constant, she had a significant MAP increase of 0.24 (with respect to the baseline group, Information Science PhD student), meaning that she was mostly picking papers at the very top of the lists. The argument about this particular user is confirmed when we see that she also had a significantly higher nDCG, precicsion@3 and precision@5, but with no strong effect on the overall precision. On the other side, a specific user, PhD student at the GSPIA program at University of Pittsburgh, found many relevant papers (more than twice the minimum requested for the task, 15) at the top and bottom of the list, resulting in a significantly higher precision than the rest.

MRR, a metric that attempts to measure how good the ranking algorithm is in locating the first relevant item at the very top, is 0.21 units significantly higher when users bookmark with the second interface. This result can either imply that users were less engaged or tired after the first subtask, so that in the second interface they bookmarked talks less selectively at the very top of the list; or it can also imply that users have become more familiar with the interface and task and, consequently, were able to find relevant papers at higher ranks (closer to the top) than when they performed the task for the first time.

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 their perception of the role of the Venn diagram.

Although MAP was higher in the controllable recommender interface, further analysis is required to generalize this result to other IR metrics (such as precision, MRR or nDCG), due to the interactivity of the controllable interface as compared to the static list of the non-controllable one.

Regressions on Subjective Metrics

This set of regressions was conducted using a linear mixed-model analysis with a Gaussian identity link. Table 15 shows three columns, the second one with a shortcut of the statement evaluated in the post-session survey (the metric) and the third 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 A**.

Table 15. Significant factors on subjective metrics collected in post-session survey.

Metric		**Significant effects [(+): positive / (-): negative]**
1	UNDERSTOOD	(+) Condition(treatment), B = 1.1, p < 0.001 (+) native speaker, B = 0.68, p = 0.003 (+) use of CN B = 0.94, p < 0.001 (+) TELCOM PhD B = 1.02, p = 0.011
2	RELEVANT	N/A
3	DIVERSE	(+) Experience in research domain B = 0.15 p = 0.04
4	INTERFACE_EASY	(-) order (second), B = -0.53, p = 0.014 (+) condition(treatment)*order(second) B = 0.97, p = 0.012
5	LOST_TRACK_TIME	(-) Experience in research domain B = -0.26, p = 0.039
6	OVERALL_SATISFIED	(-) order (second), B = -0.97, p = 0.001 (+) condition(treatment)*order(second) B = 1.09, p = 0.025
7	CONFIDENT_MISS	(-) order (second), B =-0.92 , p = 0.001 (+) condition(treatment)*order(second) B =1.19 , p = 0.012
8	USE_AGAIN	(-) order (second), B = -0.98, p = 0.001 (+) condition(treatment)*order(second) B =1.1 , p = 0.029
9	SUGGEST_COLLEAGUES	(-) order (second), B =-1.22 , p < 0.001 (+) condition(treatment)*order(second) B = 1.54, p = 0.004
10	RECSYS_NO_NEED	(-) Experience with recsys B =-0.24 , p = 0.046 (-) CMU PhD St B=-2.42, p = 0.03
These metrics only apply to the controllable recommender + order		
11	C_FEEL_CONTROL	(+) Order, B =0.79 , p = 0.014
12	C_ABIL_CONT_SATISF	N/A
13	C_ABIL_CONT_TRUST	N/A
14	C_INTEREST_EXAMINE	N/A
15	C_VENN_UNDERSTAND	(-) Gender B =-0.74 , p = 0.018
16	C_VENN_USE	(+) Trusting propensity B = 0.13, p = 0.03
17	C_VENN_TRUST	(+) Trusting propensity B = 0.19 , p = 0.004

The effect of the controllable condition (treatment) in understanding the recommendations: Using the controllable condition increases the agreement with understanding why talks were recommended in 1.1, p <0.001, keeping the other variables constant. This result is expected; only the controllable interface has explanations of the items recommended. Being a native speaker (B=0.68, p=0.003), having previous experience with CN (B=0.94, p<0.001), and being in the TELCOM PhD (B=1.02, p=0.011) compared to being an IS PhD student also positively influenced this variable.

The effect of the condition (treatment) when presented after presenting the non-controllable interface in improving the user perception on 5 out of 10 metrics: Getting quickly familiar with the interface (B = 0.97, p = 0.012), being satisfied overall with the interface (B = 1.09, p = 0.025), being confident of not missing relevant talks (B =1.19, p = 0.012) and also the user's intention of using it again (B =1.1, p = 0.029) and suggest the interface to colleagues (B = 1.54, p = 0.004)

The importance of having more experience in the research domain on two variables: Having a better perception of the diversity of the talks (B = 0.15, p = 0.04), but also a smaller perception of engagement by not feeling that the task makes them lose the track of time (B = -0.26, p = 0.039).

The experience with recsys influences the perception that Conference Navigator actually *does not need* a recommender system (B =-0.24, p = 0.046); the effect shows that experience with recommender systems significantly decreases agreement with that statement.

Gender also has a significant effect: Women perceived that the Venn diagram helped them understand why the talks were recommended significantly more than men did (B =-0.74, p = 0.018).

The positive effect of trusting propensity on perceiving the Venn diagram as useful to identify talks recommended by different methods (B = 0.13, p = 0.03), and also increasing the trust in the recommendations given the use of the Venn diagram (B = 0.19, p = 0.004).

6.2.3.2 Regression Set 2: Regressions on Controllable Interface Actions Only

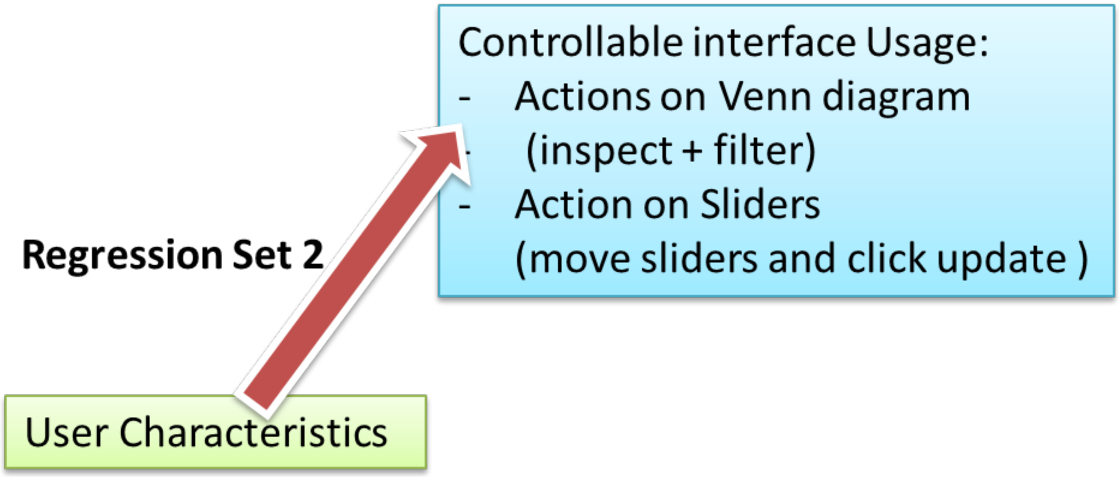

Controllable interface Usage:
- Actions on Venn diagram
(inspect + filter)
- Action on Sliders
(move sliders and click update)
Regression Set 2
User Characteristics

Figure 30. Block model of the effects and dependent variables studied in the second set of regressions.

Regressions on Actual usage of Sliders and Venn Diagram

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 16. Being a native speaker decreases the number of actions in 50% compared to non-native speakers in the Venn diagram, and 49% in the case of the sliders. Since the list of recommendations was not large (from 30 to 60 papers), native speakers could rely more on fast scanning of the talk information (title, author, paper type) at the recommendation list rather than using the visual controls.

Table 16. Significant effects of the regressions on actions over the visual widgets.

Metric	Significant effects [(+): positive / (-): negative]
Venn actions	(-) Native speaker B = -0.69, p = 0.038 / exp(-0.69) = 0.50
Sliders actions	(-) Native speaker, B = -0.67, p = 0.032 / exp(-0.67) = 0.51 (+) Trust in Recsys, B = 0.37, p = 0.013 / exp(0.37) = 1.44 (+) CN Use B = 0.6, p = 0.032 / exp(0.6) = 1.82

On the other side, an additional standardized unit of trust in recommender systems increases the number of actions on the sliders by 44%, as well as previous use of Conference Navigator increased the use of the sliders by 82% compared to those without previous experience using sliders.

6.2.3.3 Regression set 3: Regressions controlling for number of actions on the visual widgets (Sliders and Venn diagram)

To understand whether specific actions with the controllable interface had an influence on the user experience, the logarithm of number of actions on the Venn diagram and on the sliders widget were used as predictors of: (a) objective metrics, and (b) subjective metrics, but this time considering only the controllable interface. Figure 31 summarizes the analysis in a block model. The decision of using 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 with predictions and classification tasks..

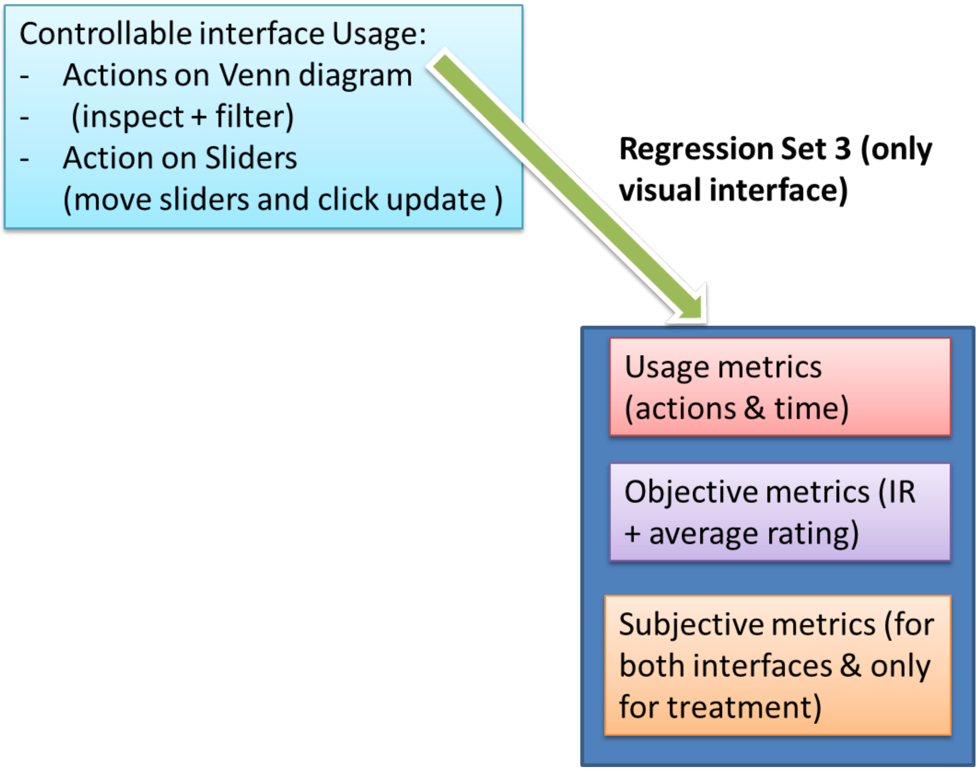

Controllable interface Usage:
- Actions on Venn diagram
- (inspect + filter)
- Action on Sliders
(move sliders and click update)
Regression Set 3 (only visual interface)
Usage metrics (actions & time)
Objective metrics (IR + average rating)
Subjective metrics (for both interfaces & only for treatment)

Figure 31. Block model of the regression on evaluation measures controlling for actions on the controllable recommender interface

The regressions performed were controlled for both variables (Venn actions and sliders actions) as predictors since the correlation among them is not significantly different than zero, rho = 0.17, *p* = 0.3

Regressions on IR metrics

Linear multiple regression was used to study the effect of the number of actions in the Venn diagrams and in 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 consider only the controllable interface (see details in the Appendix). However, we again emphasize that an extended analysis using new metrics designed to evaluate interactive sessions (rather than static lists) could reveal some effect of the actions on the Venn diagram or the sliders.

Regressions on Subjective metrics

Table 17. Significant effects found on regressions over subjective metrics applicable only to the controllable interface.

Metric		Significant effects [(+): positive / (-): negative]
These metrics only apply to the controllable recommender + order		
11	C_FEEL_CONTROL	N/A
12	C_ABIL_CONT_SATISF	N/A
13	C_ABIL_CONT_TRUST	N/A
14	C_INTEREST_EXAMINE	N/A
15	C_VENN_UNDERSTAND	(+) Venn_actions, B =0.17 , p = 0.031 (-) sliders actions B =-0.27 , p = 0.012
16	C_VENN_USE	(+) Venn_actions, B = 0.19, p = 0.004 (-) sliders actions B = -0.24, p = 0.01
17	C_VENN_TRUST	N/A

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 17. The results show a competition between the use of the Venn diagram and the use of sliders on these two statements.

On the statement “*I think the Venn diagram visualization helped me to understand why a talk was recommended*”, a unit increase in log-count usage of Venn diagram increases agreement with this statement by 0.17, whereas the opposite happens with a higher use of the Sliders. A unit increase in the log-count usage of sliders produces a decrease on agreement with the statement of 0.24 units. It is interesting, though, that there is no significant effect of the actions on the Venn diagram over the statement “I understood why the talks were recommended to me”; revealing that these two statements are not necessarily perceived similarly by users.

Regarding the statement “I think the Venn diagram visualization was useful to identify talks recommended by a specific recommendation method or by a combination of recommendation methods”, the effect is similar as with the previous statement. A more active use of the Venn diagram increased the perception that this widget was useful to identify talks recommended by one or more methods. A unit increased in the log-count of actions on the Venn diagram produces an increase of 0.17 units in agreement with the aforementioned statement, while the opposite happens when people use the sliders more. A unit increase in the log-count usage of sliders decreased the agreement with the statement in 0.24 units.

In summary, the results on users’ perception suggest that the sliders and the Venn diagram were not completely complementary and they rather compete on certain aspects of explainability on the interface.

6.2.3.4 Summary of Statistical analysis

The three sets of regression analysis on metrics of different dimensions (usage, objective and subjective) provide answers to the research questions stated on this study. The next subsections summarize implications on the results to the four research questions.

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) indicate that the users engaged with the controllable recommender interface. Although one question related to focused attention (*…I lost track of time …*) did not show an effect in the controllable interface, the effect was large enough to be significant on the other three dimensions when participants used it as second interface, after performing the bookmarking task with the non-controllable interface. Under this condition, users significantly agreed that the controllable user interface was easy to learn, that they would use it again and that they would suggest it to colleagues.

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 only two important metrics: Mean Average Precision (MAP), and understandability of the interface. The first one is an objective metric used frequently in IR, MAP, and it shows that the controllable interface does a better job at ranking relevant items to the top. This result is important but the fact that a single list in the non-controllable interface is compared to a set of dynamic lists in the controllable recommender suggests that an additional analysis using recently introduced IR metrics for interactive user sessions would give stronger support for this result. The other metric is understandability, a subjective metric evaluated in the post-session survey. It shows that the design of the controllable interface actually triggers a better perception in users about understanding what is being recommended compared to the non-controllable one.

Another five subjective metrics show a positive effect of the treatment (the controllable interface), but the effect was stronger 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.

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 her research domain had significant effects on users’ engagement.

On the first set of regressions we identify that users are more likely to explore the talks using the traditional actions (checking the abstract rather than exploring the new features) if they have previous experience with the system, Conference Navigator. 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, in this section 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 very important because both characteristics had different effects on the user behavior in this study: high trusting propensity makes people have a better perception of a new feature without a clear visual affordance like the Venn diagram, and high trust on recommenders makes people more willing to use the sliders, which is a more traditional and widespread visual 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. It would be interesting to see whether this behavior is replicated when the list of recommended items is in the order of hundreds or thousands, unlike this study, which showed lists of talks of no more than 60 items.

The third most important user characteristic influencing engagement was the expertise of the user in her research domain. Participants with higher levels of expertise in their domain felt less immersed in the use of the interface, since they did not agree with the phrase "*I lost track of time while I was using the recommender interface.*"

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

User characteristics had important effects on the user experience in the recommender system investigated. The user engagement dimension is part of the user experience, and the analysis of RQ3 showed the effect of user characteristics like trusting propensity, experience on her research domain and previous use of Conference Navigator. A high trusting propensity led to a positive perception of the Venn diagram as a tool that increased trust on the recommendation list and helped to identify the fusion between recommender methods. However, in terms of the perception of understanding why a talk was recommended, the most important effect was gender: males agreed less than females that the Venn diagram helps them understand why a talk was recommended.

Another important variable for the user experience was being native speaker. In RQ3, it was already shown that native speakers performed fewer actions on the sliders and on the Venn diagram. In terms of understandability, these participants had better perception than non-native English speakers in understanding why the talks where recommended.

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. 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*.

Finally, the experience with recommender systems had a significant effect on perceiving that Conference Navigator actually needs a recommender system. This same user characteristic was associated with less exploration of the talks’ abstract, so probably these users understand the benefit of a recommender system in reducing information overload and perceive it similarly in this study.

6.2.4 Behavioral Analysis

To acquire a better understanding of how users utilized the controllable interface compared to a more traditional statics set of recommendations, an exploratory behavioral analysis was conducted and additional hypotheses were tested. In this section, the first subsection addresses the exploratory analysis to describe how subjects used the several features available in the controllable user interface.

In order to better understand the following sections Table 18, which was already presented in section 3.2.1.4, presents the actions or clicks that users could make on each interface:

Table 18. List of actions tracked in the recommender interfaces (Controllable and non-controllable).

Action	Description	Visual Widget	Controllable,	Non-Controllable
clickRetrieveList	Retrieve initial list of recommendations	Recommender interface	X	X
scheduling	Bookmark a talk	Talk	X	X
unscheduling	Remove bookmark	Talk	X	X
seeMore	Expand list or recommendations	Recommender List	X	X
clickOpenAbstract	Open abstract of talk	Talk	X	X
clickCloseAbstract	Close abstract of talk	Talk	X	X
changeSlider[N]	Change weight of method N	Slider Widget	X	
clickUpdateList	Update recommendation list	Slider Widget	X	
hoverMethod[N]Explain	Show explanation of method N	Slider Widget	X	
hoverCircle[N]	Mouse over circle (talk) on the subarea (method) N	Venn diagram	X	
clickEllipse[N]	Click Ellipse (Venn diagram) to filter list by method N	Venn diagram, list of talks	X	

6.2.4.1 Amount of subjects using each feature

Figure 32 shows the amount of subjects that used each action available, with the x-axis showing the number of users and the y-axis the action names. The color red is used for counts of subjects in the on-controllable 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 subject in the study is 40, and, as a within subjects study, all of them used both conditions: non-controllable interface (0 and red in the plot) and controllable interface (1 and green in the plot). Under both interfaces, we 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).

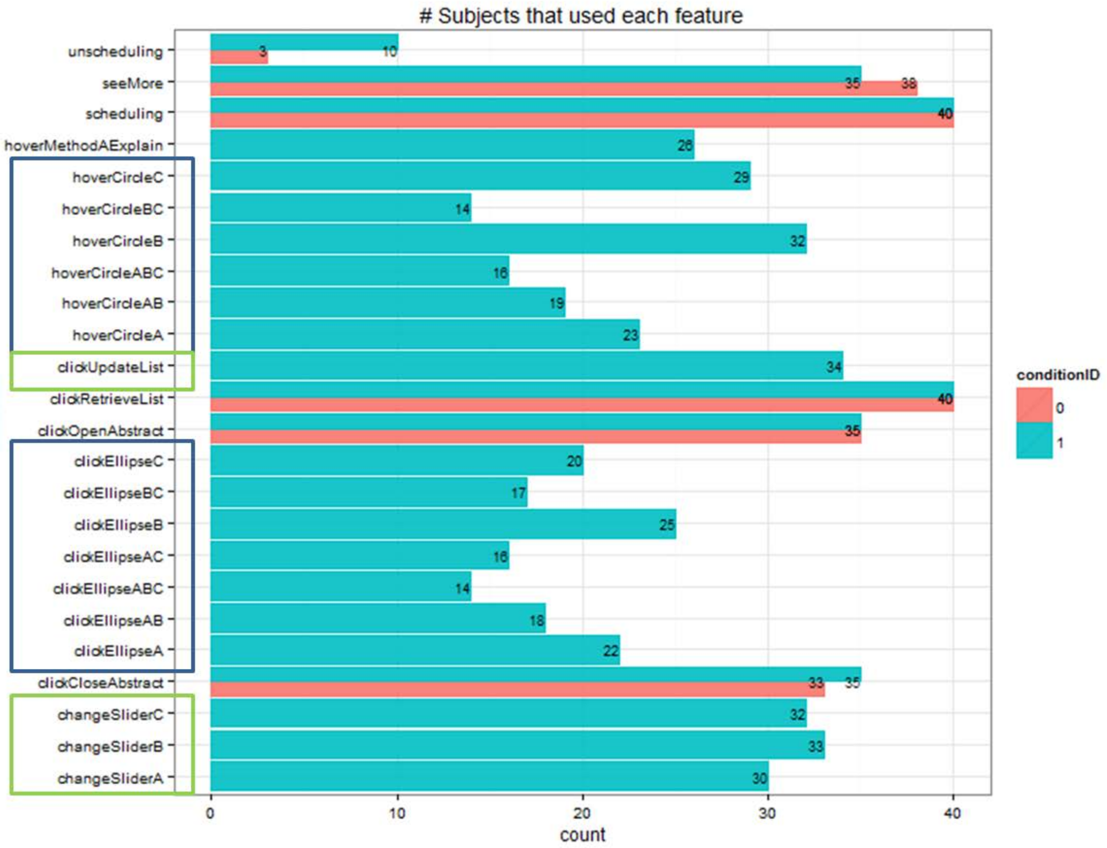


Figure 32. Plot of the amount of subjects that used each action available in the recommender interfaces.

95% of the subjects (38) in the non-controllable interface expanded the list of recommendations with the action *SeeMore*, compared to 87.5% of subjects (35) on the controllable interface. 87.5% of people (35) 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 (33) who *closed the abstract* in the non-controllable interface.

The first relevant difference between the two interfaces is in the number of people who unscheduled papers, i.e., that changed their mind after bookmarking a paper and removed them from their list of relevant papers, with 25% of people (10) in the controllable interface and only 7.5% of people (3) in the non-controllable one. This can be interpreted as a sign of greater engagement and willing to interact with the interface.

With respect to actions only available in the controllable interface, more people used the sliders than the Venn diagram features. Figure 32 shows that of actions available in the sliders, at least 75% people (30) used the *changeSliderA* –to change the weight of the method based on talk popularity- and at the top 82.5% of people (33) clicked in the button *ClickUpdateList* to reorder the list of recommendations. With respect to the Venn diagram, the action that most people tried was *hoverCircleB*, with 80% of the subjects (32), which corresponds to positioning the mouse over circles on the content-based recommender to display its title on a floating dialog. The one that less people tried was *clickEllipseABC*, with 35% of people (14) trying it, which corresponds to clicking on the intersection area of the three ellipses that filters the papers in the list recommended by the three methods.

6.2.4.2 Average number of actions per each Action Type

In the next plot we examine the average number of actions per user for each type of action in the controllable interface. Figure 33 provides these results and we can see that the most frequent action io average is *clickOpenAbstract* with 17.3 clicks per user, and the second one is *scheduling*, which averages 15.8 clicks per user. This is very close to the average number of bookmarked papers, and is not exactly the same because of some subjects that change their mind and also *unschedule* some talks.

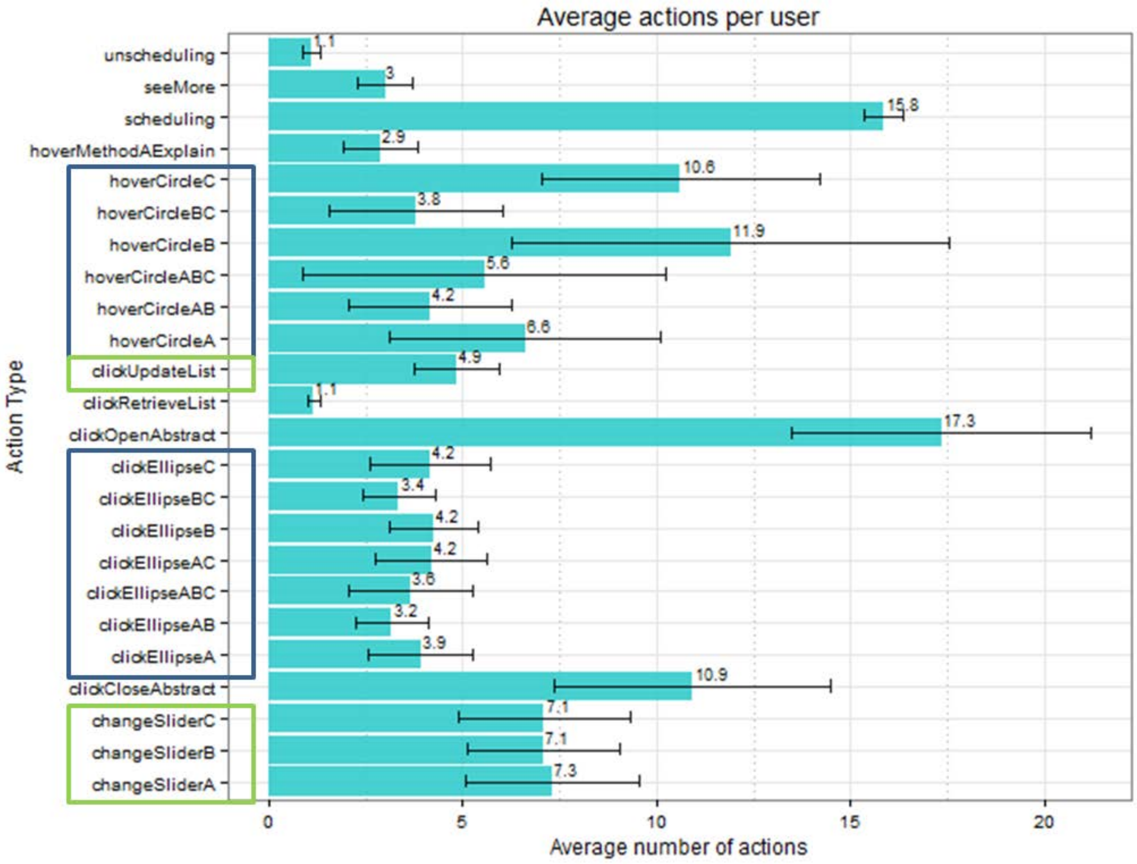


Figure 33. Average number of actions per user (x axis) at each type of action (y axis), in the controllable recommender interface.

Among the sliders actions, which are highlighted in Figure 33 with a green frame around their name in the y axis, users' change them in average close to 7 times each, without much difference across the methods (*changeSliderA*, *changeSliderB*, *changeSliderC*). The action to re-sort the recommended list, *clickUpdateList*, was carried out 4.9 times per user in average.

Regarding the actions in the Venn diagram, among the hovering actions to inspect the title of papers represented as little circles, the most frequent action was *hoverCircleB*–articles recommended by the content-based method-11.9 times, then *hoverCircleC*–articles recommended by the popularity of the authors- 0.6 times and third *hoverCircleA*–talks recommended based on their popularity-only 6.6 times. Interestingly, there is not much difference between this action and *hoverCircleABC*, 5.6 times, which are hovering actions over talks recommended by the 3 aforementioned methods. Now, with respect to *clickEllipse* actions that are a competing alternative to the *clickUpdateList* because it filters the recommended list of talks, the top three most used filters are *clickEllipseB*–to filter content-based recommended talks– 4.2 times, *clickEllipseC*–to filter talks recommended by the authors' popularity-4.2 times, and then *clickEllipseAC*–the combination of popular papers with papers with popular authors-in third place with 4.2 times, leaving *clickEllipseA* in fourth place with 3.9 clicks on average. Actually, the difference in terms of actions per user among these filters is not very large, so a more interesting question is which one is used more frequently to bookmark papers, and that is the analysis of the next subsection.

6.2.4.3 Effectiveness of each filtering method for bookmarking

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 and then clicking on *Update Recommendation List*. Table 19 shows the amount and percentage of bookmarks done using each of the three options just described. There were a total of 616 bookmarks made in the controllable recommender interface, and more than half of them (58,44%) were done after updating the list controlling through the sliders. This proportion is significantly different from 50%, after performing a one-sample test of proportions,χ^2=17.22, $p < 0.001$. Moreover, using a one-sample test of proportions we found that the 28.08% of bookmarks performed after filtering with the Venn diagram is significantly different than the 13.47% of talks bookmarked without any previous filter, χ^2=100.33, $p < 0.001$.

Table 19. Distribution of bookmarks by filter in the controllable interface

Without Filters	Using sliders	Using Venn diagram
83 (13,47%)	360 (58.44%)	173 (28.08%)

In terms of detailed filters, Figure 35 shows a summary of the detailed filters sorted by number of bookmarks scheduled. To understand the filter labels, 2 conventions are used:

1) Venn filters are encoded with letters and pipes, as shown in Figure 34. The code |B|||| means that only the area of papers recommended by the content-based recommender was active. A code |B|C||| is different from |||||BC|, in which in the first case the user selected papers recommended only by method B and only by method C, and in the second case the user selected papers recommended by methods B and C.

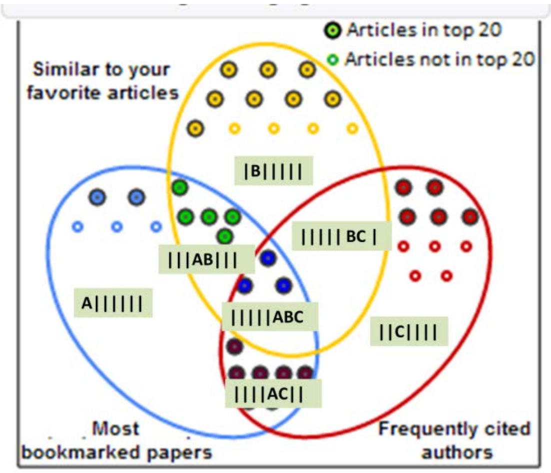


Figure 34. Codes used to identify the different areas to filter in the Venn diagram.

2) Sliders are encoded based on the weights given to methods A (papers frequently bookmarked), B (content-based recommender), and C (paper with authors frequently cited), If detailed filter *weights 0,1,0*, this means method A: 0, method B:1, and method C: 0, i.e., the filter will show at the top papers recommended only by method B.

Figure 35 shows all of the filter used to bookmark at least 5 papers, totaling 555 (around 90% of all the bookmarked talks). At the top we see *D_weights_0.4,0.6,0.4* with 86 bookmarked talks. This filter is actually the default filter, i.e., when the user bookmarked without using either the sliders or the Venn diagram in order to do so. In second place we see |*B*||||, which is the method B (content-based recommender) after the Venn diagram filter, with 55 bookmarks. The third one is *weights 0,1,0*, which is exactly the same as the |*B*|||| but using the sliders instead.

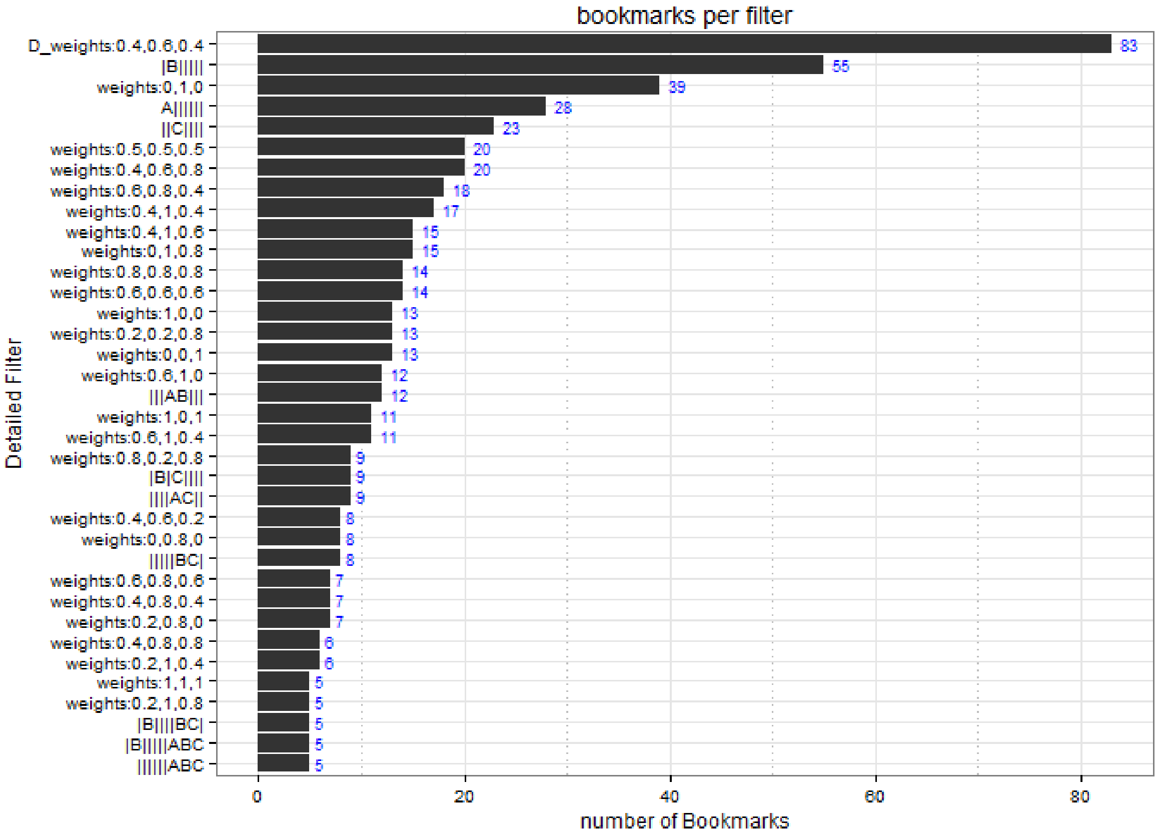


Figure 35. Number of talks bookmarked using each detailed filter.

The plot in Figure 35 expresses two important concepts. First, that although most talks overall were bookmarked after using the sliders, the most popular filter in detail was actually the content-based recommender ellipse of the Venn diagram; and among the top 5 filters we also see the popular talks (A||||||) and popular authors filter (||*C*||||) in third and fourth places. Second, the top two filters are the content-based recommender, since |*B*||||| and *weights 0,1,0* are equivalent, but the first one uses the Venn diagram and the second uses the sliders. After those top 4 filters, we see more use of blending the weights of different methods using the sliders.

6.2.4.4 Ranking and Ratings at each filtering method in both interfaces

Two metrics that can give more information about the performance of the filters compared to not using them –in the controllable and the non-controllable interface– is the average ranking of the paper in the list at the moment of being bookmarked, and the average rating received by the papers recommended by different methods. In this subsection these results are presented and discussed.

Table 20 shows the average ranking of the talks at the moment of being bookmarked, classified by different filters. The lowest average ranking of the Venn diagram's bookmarked item (5.82) is expected since the resultant filtered lists by this method are shorter than the lists loaded by default (30 items) and also shorter than the lists produced after the reordering with the sliders (30 items too). Although the average rank of the bookmarked items in the non-controllable interface is the highest (21.35), what is more interesting is the second highest: the sliders' list, whose average rank is 19.84, even higher than the default list, with 17.19. This might explain why among the top 5 most used filtering methods in the recommender interface are Venn diagram filters, since the users are really decreasing the potential list size from which to choose a relevant item.

Table 20. Average ranking of the talk at the moment of being bookmarked, classified by filter.

Controllable interface			Non-controllable interface
Default List	Sliders Lists	Venn diagram lists	Default Lists
17.19	19.84	5.82	21.35

Finally, an interesting metric to analyze –given that there was no significant difference in average user rating among the interfaces- is the average rating of papers classified by the method they were recommended, and comparing these rating between both interfaces.

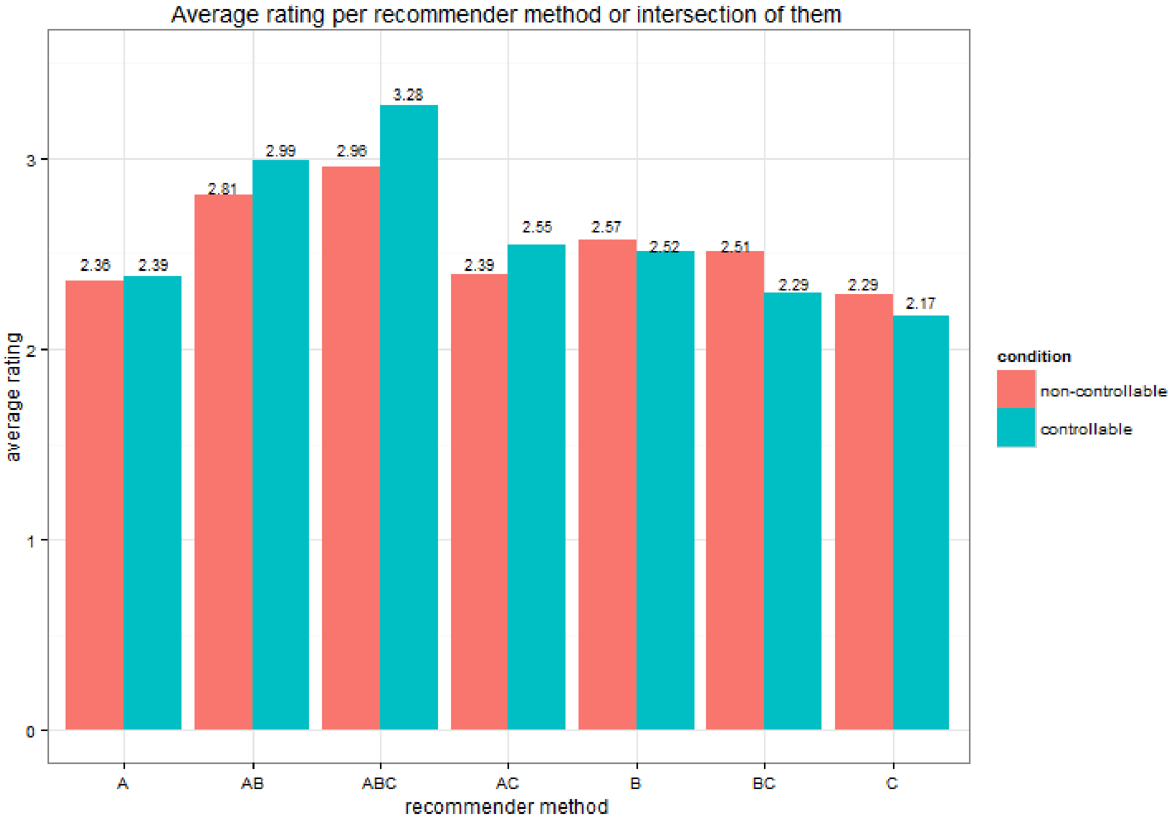


Figure 36. Average rating per recommender method (or intersection of them) under the non-controllable and controllable interfaces.

Figure 36 shows a plot of average ratings of different recommendation methods (or intersections of them) separated by interface (non-controllable and controllable conditions). Let’s remember that method *A* is popularity based on bookmarks, method *B* is the content-based recommender, and method *C* is popularity based on authors’ citations. It is interesting that both distributions of ratings are similar between both interfaces, although the users in the controllable interface were able to actually see that certain papers were recommended by more than one method, unlike the users of the non-controllable interface. If we compare the average user rating between the two interfaces for each method, we don't find significant differences. For instance, the largest difference between interfaces is observed in the method ABC for the controllable (M=3.28, S.E.=0.27) and the non-controllable interface (M=2.96, S.E.=0.29), and is not significantly different, using a related-samples Wilcoxon Signed rank test, p = 2.93. This result means that the methods or fusions of methods perceived as more relevant are the same between interfaces. However, within the same interface, the relative perception of relevance among fusion of methods and single methods was found different between interfaces. More specifically, within the non-controllable interface, the average user rating of talks bookmarked with fusion of methods ABC 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 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.

In summary, the analysis indicates that fusing several recommender methods can produce a better perception of relevancy (for instance, the fusion of ABC is better than only C in both interfaces), and this effect might not be significant when we compare the same method in different interfaces (e.g., no significant difference between ABC in the controllable and the non-controllable interface). Moreover, highlighting the fusion of methods utilizing specific visual components (colors and the Venn diagram) can have an effect on the relative user' perception of relevance between methods within the same interface (e.g., comparing the difference between ABC and A within each interface).

The two results just indicated show that the inherent relevancy provided by each method within each interface is rather equal, however, if we compare between interfaces there seems to be an effect of the interface highlighting which combination of methods is more or less relevant.

6.2.5 Post-Study survey

At the end of the study session, each subject answered a post-survey in order to collect the user perception when comparing the recommender interfaces: the static list and the visual controllable one. This survey was composed of 6 questions, the first five were multiple- choice questions and the results are summarized in Table 21. A qualitative analysis of the user comments from question 6 is presented in the next subsection.

Questions 1 and 2 asked participants about their preferred interface and which one they would suggest to permanently implement in Conference Navigator. The results of both questions show a clear preference for the visual controllable interface. For the first question 36 out of 40 (90%) preferred the visual controllable recommender and 4 out of 40 (10%) subjects 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.

Table 21. Results of the post-survey. Numbers indicate the amount of subjects that chose the indicated option.

1. Which one of the interfaces did you like/prefer most?			
The static list of recommendations	The visual controllable recommender	I liked both of them	I didn't like any of them
0	36	4	0
2. Which of the interfaces would you suggest to implement permanently in Conference Navigator?			
The static list of recommendations	The visual controllable recommender	I wouldn't suggest to implement any	I would suggest to implement both
1	33	1	5
3. Which of the interfaces did you feel that required more effort in order to find relevant articles?			
The static list of recommendations	The visual controllable recommender	Both required more or less the same level of effort	I cannot tell which one required more effort
24	7	5	4

	(1) I don't like it at all	(2)	(3) I don't know	(4)	(5) I really like it
4. Overall how would you rate the static list recommendations interface? (2.875)	4	8	17	11	0
5. Overall how would you rate the visual controllable recommendation interface? (4.425)	0	1	0	20	19

Although there is 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%, $\chi^2 = 28.36$, $p < 0.001$.

6.2.5.1 User comments of Post-study survey

The last question in the post-study survey asked subjects whether they could elaborate why they preferred one interface over the other. Interesting insights were obtained and this section summarizes them.

6.2.5.2 Positive comments on the visual interface

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 characteristics:

"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."

Another related comment on how helpful the Venn diagram was to find relevant talks:

“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.”

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 affect controllability in the precious 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. In particular, one user said that he preferred the sliders because he was familiar with a similar control in another system:

“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.”

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 used 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 used 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."

Visual interface is easy to learn and is fun!

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.

One person expressed in his comments that using the controllable recommender made the task less boring:

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

and another that she found the interaction engaging:

“the controllable recommender is interactive and engages me more. Gives me more control”

6.2.5.3 Critical comments in the visual interface

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 2 to zero 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.

With these comments in mind, there is a potential for personalization or at least for customization, in letting user view some of the visual widgets while they interact with the interface.

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 papers, 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. The fact that it is not easy to asses th relevancy of these items (papers) makes this tool useful for several users, but it would be interesting to assess if its usability increases under a scenario where users have to find a few relevant items among thousands or millions of possible options.

Disconnection between sliders and Venn diagram

One user commented that there was a disconnect 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 more clear to users how it works to avoid confusion or lose of trust in the interface.

Allow me to remove things I dislike

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

Bookmarked paper should have a different color

One subject complained that the Venn diagram colored the bookmarked items with grey: *"I like the visual controllable recommender. Interface more. However, it's not my ideal interface. One suggestion for Venn diagram is that the current gray circles that show bookmarked articles should have another color. Gray conveys something inactive, something in background, while bookmark indicates foreground and being active."*

6.3 SUMMARY AND DISCUSSION

This study has addressed the four research questions of this dissertation. The analyses conducted on this study comprised three different kinds of metrics: usage (actions and time spent), objective (rating and IR metrics), and subjective (dimensions evaluated through surveys). Regarding research questions 1 and 2, the results show a positive effect on user engagement and user experience with the controllable recommender interface, but the effect is stronger when it was presented after the non-controllable one. The number of bookmarked talks and the mean average user rating was not different between interfaces, but the mean average precision (MAP), an IR metric, and several subjective metrics from the post-session survey supported a positive effect of the treatment (the controllable interface) on the user engagement and experience.

User characteristics also had a significant effect. With respect to usage metrics, having previous experience with Conference Navigator increased the number of actions to explore talks using the abstracts, whereas having experience with recommender systems decreased the number of actions of the same metrics. Being a native English speaker decreased the number of actions with the visual widgets (the sliders and the Venn diagram), but they reported a significant agreement in terms of understanding why the talks were recommended. An interesting competition relation was discovered with respect to number of actions on the Venn diagram and actions on the sliders under the controllable interface. The first, as expected, increased the perception that the Venn diagram was useful to understanding the fusion between recommenders and also the feeling of not missing relevant talks, while a higher number of actions on the sliders widget produced the opposite effect. Other characteristics such as experience with research domain and trusting propensity also showed significant effects.

The analysis of the post-study survey and comments of the participants showed a more biased perception in favor of the controllable interface, the favorite of the users in several dimensions. There were many positive comments; the most enthusiastic particularly praised the Venn diagram for providing a clear and straight forward filtering method considering a single or a multiple combination of recommender methods. On the other side, a recurrent comment was that the controllable interface would work better with a larger number of recommendations. Some users considered the interface too complex for such a small number of items (lists of up to 60 items) that could be scanned and scrolled instead of learning the filtering widgets.

6.3.1 Path analysis: User Characteristics and Actions on the Controllable Interface

An interesting result was found by conducting two sets of regression analyses on the controllable interface. With the first set, controlling for user characteristics to predict number of actions with the sliders and with the Venn diagram, we learned that being a native English speaker reduced the number of both types of actions, but having trust in traditional recommenders and with Conference Navigator increased the number of actions with the sliders. No user characteristic was found significant for increasing the number of actions on the Venn diagram. The second set of regression predicted the agreement on the post-session survey controlling for the log of the number of actions on sliders and on the Venn diagram. The result shows a significant but opposite effect of the predictors. More actions in the Venn diagram increased the perception of Venn diagram as a tool to explain fusion between methods and to increase trust of the recommendation list. The opposite effect is shown by the increased number of actions of the sliders, decreasing the positive perception over the Venn diagram. We think that this result reflects two types of users: those willing to try newer features and those who prefer to use more familiar tools. Users more familiar with the system (Conference Navigator) and those who have followed recommendations in the past are more likely to use sliders, a more familiar way to control a recommended list then the Venn diagram.

6.3.2 IR Metrics in Static and Interactive Sessions

Although MAP was found to be significant higher in the controllable recommender interface, further analysis is needed to generalize this result to this or any other IR metrics (such as precision, MRR or nDCG), since the interactivity of the controllable interface prevents straight-forward comparisons between both conditions. The non-controllable interface requires calculating a single metric per user, given a unique list of recommended items; however, the controllable condition requires calculating the metric several times over the different lists that the user generates by interacting with the interface. It is not clear that this approach results in a fair comparison, and only recently have new IR metrics been created to evaluate the performance of an interactive session rather than a single list, such as the session-based DCG (sDCG) proposed by (Järvelin, Price, Delcambre, & Nielsen, 2008).

6.3.3 Insights from Post-study Survey: List Length and Negative User Intervention

Important feedback was given the users at the end of the post-study survey. Many of them praised the visual features, interaction design and transparency of the controllable interface, but some of them highlighted important observations. The most important is that the system is not very useful if the list of recommendation is not long, which makes the visual features unnecessarily complex because the items can be just scanned one by one. Although the great majority found the system easy to learn, some people felt that having to learn how to use the sliders and the Venn diagram, in addition to completing the task of bookmarking papers, was too demanding. Here there is a potential for adaptation: the interface can allow the user to remove or temporarily hide some of the features to avoid excessive cognitive load when she feels it. Another point repeated from the feedback of the CSCW study that a user expressed is that he would like to tell the systems which items were disliked and remove them. This comment highlights the need to provided the user the means to express not only his positive opinion but also his negative one, which can affect the trust that she has in the system.

6.3.4 Behavioral Analysis: Impact of Filtering Mechanisms

Finally, the exploratory behavioral analysis complements the analysis of objective and subjective metrics, showing that subjects made active use of the system (58% of the talks were bookmarked after using the sliders and 28% withan active Venn diagram filter), and that they had clear preferences for specific filtering methods (three of the top 5 filtering methods were from the Venn diagram) and recommendation approaches (the content-based method accounted for the largest portion of bookmarks). The analysis of the average ratings on both interfaces by different recommendation methods shows that the best and worst rated methods are the same in both interfaces, independent of visual widgets, but the relative difference of ratings suggests an effect of the interface favoring popular bookmarked papers in the controllable interface and papers recommended by the popularity of the authors in the non-controllable one.

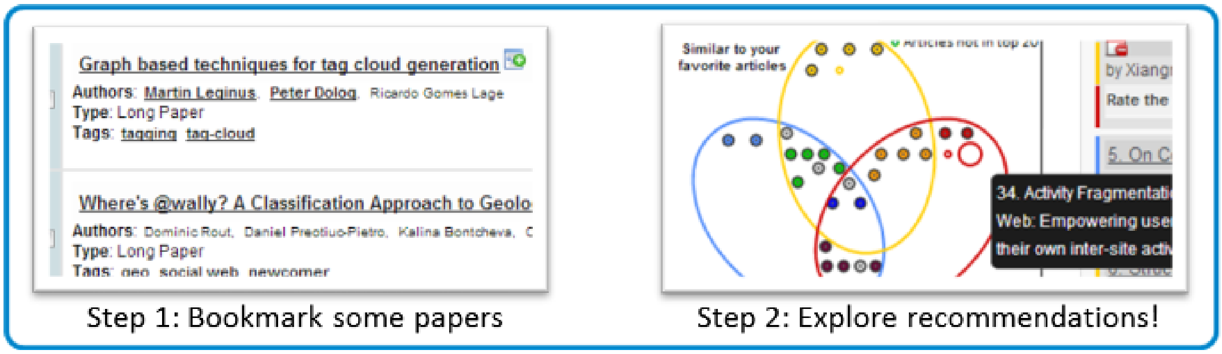
7.0 CASE STUDY: CONTROLLABLE RECOMMENDER IN A REALISTIC SETTING

In order to compare how the controllable recommender interface is used in a realistic setting versus a laboratory setting we tracked and analyzed its usage in two conferences: ACM Hypertext 2013 (May 1-3 2013) and UMAP 2013 (June 10-14 2013). This chapter summarizes the usage of the recommender interface during both conferences, presents the analysis of metrics to compare the users' behavior between the realistic and the laboratory setting, and it summarizes the results of a survey conducted after the aforementioned conferences finished.

Although the Study 1 was previously conducted in the realistic setting of CSCW, two important factors justified a new field study. For one thing, the preference elicitation in CSCW was rather unnatural compared to other conferences supported in Conference Navigator (three elicitation steps not based on bookmarks before obtaining recommendations). Second, the Venn diagram filtering, which accounted for almost 30% of the bookmarks in Study2, was still not fully implemented in Study 1.

The usage data presented in this section was logged starting two weeks before each conference took place and finishing at the end of June. To promote the system, an e-mail was sent to users of Conference Navigator that attended to previous versions of these conferences announcing the availability of the controllable recommender. Unlike the laboratory use study, conference users were not given any prior training on how to use the recommender interface.

Conference Navigator featured a promotion message to encourage usage of the recommender. The users of both conferences, HT and UMAP, saw the promotion message shown in Figure 37 in the homepage after they logged in.



Graph based techniques for tag cloud generation
Authors: Martin Leginus, Peter Dolog, Ricardo Gomes Lage
Type: Long Paper
Tags: tagging tag-cloud

Where's @wally? A Classification Approach to Geolo
Authors: Dominic Rout, Daniel Preotiuc-Pietro, Kalina Bontcheva, C
Type: Long Paper
Tags: geo social web newcomer

Step 1: Bookmark some papers
Similar to your favorite articles
Articles not in top 20
by Xiang
Rate the
5. On C
34. Activity Fragmentati
Web: Empowering use
their own inter-site acti
Step 2: Explore recommendations!

Figure 37. Promotion message for paper recommendation displayed in the conference home page. Every user saw this message after logging into the system.

The users were allowed to explore the system without restriction. They could use all the features, their actions were tracked in our database and an e-mail was sent a week after the conference finished in order to ask for users' feedback regarding their experience using the recommender.

7.1 RESULTS

7.1.1 General Usage Metrics in HT and UMAP 2013

Table 22 shows usage statistics in both conferences. The number of people using Conference Navigator during HT 2013 is smaller (51) than those in UMAP 2013 (95), but the conversion rate for the recommender interface (the percentage of users logged in the system that accessed the recommender interface) is similar between both interfaces, around 50%. 24 people did a total 306 bookmarking actions during HT 2013, with a 7.9% (24) of the bookmarks done by 3 people using the recommender interface. In UMAP 2013, the bookmarking actions summed up to 731 done by 56 people, where 14.1% (103) of the bookmarks were done by 14 users using the recommender interface. The amount of actions that users performed using the controllable interface in HT 2013 was 266, done by 27 users. In UMAP, 823 actions were performed on the recommender interface by 50 users. In terms of user feedback, 26 ratings were provided by 2 users in HT, with a mean average rating of 3.16, whereas in UMAP 8 users provided 86 ratings to their recommendations, with a mean average user rating of 3.62.

Table 22. Usage statistics of Conference Navigator during HT and UMAP 2013.

	HT 2013	UMAP 2013
Authenticated users in CN	51	95
Users who bookmarked in CN	24	56
Users who visited the recommender	27	50
Users who bookmarked papers using the recommender	3	14
Number of bookmarks in CN	306	371
Number of bookmarks using the recommender	24	103
Number of ratings in the recommender	26	86
Mean average user rating	3.16	3.62
Number of actions in the recommender interface	266	823

7.1.1.1 HT 2013 User Interaction

Figure 38 shows the average number of user actions for each type of action. The blue number in parenthesis shows the amount of people who did the respective action. In HT 2013, we see that, in average, people were more likely to rate than to schedule papers. This is probably a behavior similar to the one shown in CSCW, where people provide ratings as a form of user intervention to receive better recommendations.

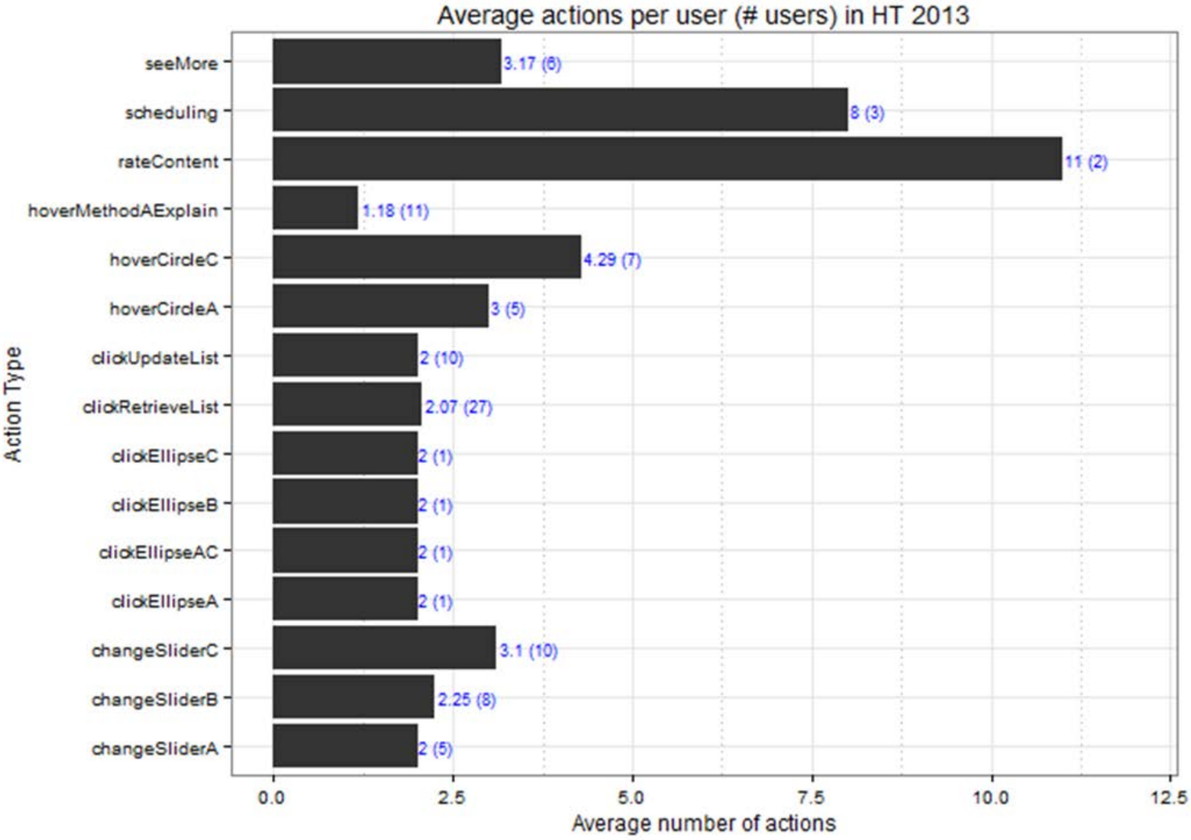


Figure 38. Average number of user actions per type of action in HT 2013. The number in parenthesis shows the amount of users that performed the respective action indicated in the y-axis.

In Figure 38 we can see that only one person tried the actions of type *clickEllipse* in order to filter the list of talks using the Venn diagram. More users tried the sliders with 10 people that clicked on the button *clickUpdateList* with a higher prevalence of the method C (action *changeSliderC*, papers recommended by the popularity of their authors) with 3.1 actions in average by 10 people, compared to 2.25 actions in average of method B (action *changeSliderB*, content-based recommender) performed by 8 people, and 2 actions in average performed by 5 people in method A (action *changeSliderA*, papers recommended by number of bookmarks). Recalling the use of the Venn diagram in Study 2, we see that participants used it differently during the conference. Only 1 person tried the *clickEllipse* action over the Venn diagram, and 7 people performed the *hoverCircle* action (mouse over a circle representing a paper that displays a floating dialog with the paper title). Among the visual widgets actions, this was the one with the highest number of actions per user, and particularly over the method C (4.29 times in average by 7 users) and over the method A (3 times in average by 5 users).

One explanation for the difference in users' behavior in the HT and UMAP 2013 conferences with respect to studies 1 and 2 in terms of which recommendation methods generate more attention from the users, is that in the HT conference many users faced *cold-start*, i.e., they did not have bookmarks from the present or previous conferences and then the Venn diagram did not show any content-based recommendations (method B). Another important difference in this real setting with respect to the laboratory study is that here the list of recommended talks based on popularity is dynamic: it changed over time based on users' bookmarks. In the laboratory study, all the users saw the same set of popularity-based recommendations. Only recommendations generated with method C do not change over time because the papers are ranked based on the number of citations of the authors previous to the conference, giving it more stability in the amount of recommendations and the papers recommended then the other two recommendation methods.

7.1.1.2 UMAP 2013 User Interaction

Figure 39 shows the distribution of average user actions in UMAP 2013. Although the user behavior shown in Figure 39 for UMAP 2013 departs from the one seen in the laboratory setting of Study 2 (Figure 33), it is closer than the user behavior displayed in HT 2013.

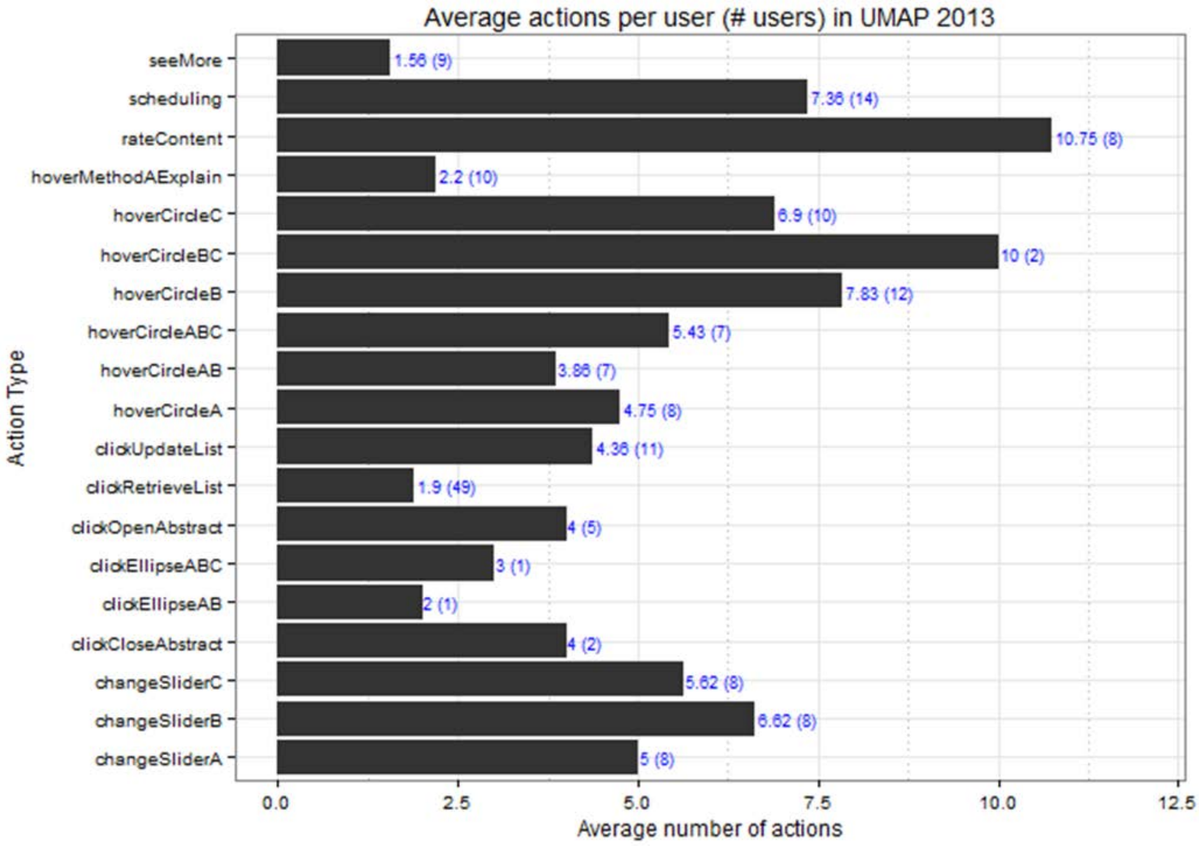


Figure 39. Average number of user actions per type of action in UMAP 2013. The number in parenthesis shows the amount of users that performed the respective action indicated in the y-axis.

The main difference between UMAP 2013 users' behavior and the laboratory study is that only 1 person used the Venn diagram filtering (*clickEllipse*), which is completely unexpected given the strong evidence in study 2 of its helpfulness to filter talks with several contexts of relevance. We believe that users simply were not aware of this feature. Users were actually aware of the Venn diagram and they used it, since they inspected talks using the *hoverCircle* action. This action allowed them to examine papers recommended by one or more methods by moving their mouse over the circles in the Venn diagram. Proportionally, the usage distribution of different *hoverCircle* actions –in terms of which recommender method attracts more user attention– resembles the one in the laboratory study with the exception of the *hoverCircleBC* action, which is among the lowest in Study 2 (3.8 times in average) and here is the top one (10 times in average), but only used by 2 people. We observe in UMAP 2013 data a bigger interest for examining papers recommended by method B (content-based recommendations) and its intersections compared to HT 2013. The reason might be that several UMAP attendees are also CN users from previous conferences, so they actually had received content-based recommendations from the first time they accessed the interface, unlike HT 2013 users. In addition to the Venn diagram people also used the sliders to control their recommendation lists, though they show a more conservative behavior by performing less *changSliderN* and *clickUpdateList* actions than in Study 2 with only around 5 actions per user instead of 7 actions in average. Finally, another important difference with Study 2 is that people are less likely to examine the abstracts (action *clickOpenAbstract*) in this real setting: only 4 times in average by 5 users compared to 17.3 times in average in the study 2. In the UMAP 2013 conference, just as in HT 2013, people were more likely to rate more papers (10.75 times by 8 users) than to bookmark them (7.36 times by 14 users) in average.

7.1.1.3 Summary of General Usage Metrics

By examining the average usage actions distribution it is possible to observe that users effectively utilized the most prominent visual features of the controllable recommender interface, sliders to update the recommended list and the Venn diagram to inspect the talks. Furthermore, there is a distinguishable difference in users' behavior between the laboratory setting: with HT and UMAP: the lack of user actions of the type *clickEllipse* in the realistic setting, which allowed users to filter the recommendation list by clicking on the white area of ellipses and intersections of them in the Venn diagram. This action was praised by many users in the laboratory setting of

Study 2, accounting for almost 30% of the total number of bookmarks after clicking on the Venn diagram ellipses. Users considered that filtering papers recommended by more than one method through the Venn diagram helped them to find relevant papers. It is possible that this action did not have a simple affordance, and the fact that users did not have a prior training on the interface could have contributed to the behavior of ignoring it.

7.1.2 Comparing Laboratory results with UMAP 2013

In this section three usage metrics are studied to understand the user behavior in a realistic setting. The metrics compare the user behavior between the controlled setting of Study 2 and the realistic setting of a conference using UMAP 2013 data, since it is more abundant than HT 2013 and then it allows conducting statistical tests.

7.1.2.1 Proportion of people who used the sliders

A chi-square test was conducted to study if the expected proportion of people using the sliders based on the results of Study 2 (30 out of 40), is significantly different than the proportion exhibited in the UMAP 2013 conference (11 out of 27). The test shows that the proportions are not significantly different (chi-square = 1.45, df=1), p=0.22. On this way, we conclude that the proportion of people that use the sliders in the controlled laboratory setting and in the real setting of UMAP 2013 is not significantly different.

7.1.2.2 Proportion of bookmarks by filtering method

A chi-square test was conducted to study if the expected proportion of bookmarks using the sliders based on the results of the laboratory study (360 out of 616), is significantly different than the proportion exhibited in the UMAP 2013 conference (33 out of 103). The test shows that the proportions are significantly different (chi-square = 7.77, df=1), p=0.005. On this way we reject the null hypothesis and we conclude that the proportion of papers bookmarked after using the sliders in the controlled laboratory setting (58%) is significantly higher than in the real setting of UMAP 2013 (32%).

7.1.2.3 Proportion of papers that were inspected (*hoverCircle* action) and bookmarked

A chi-square test was conducted to test differences in the proportion of papers that were inspected in the Venn diagram interface and then bookmarked, comparing between the laboratory study and UMAP 2013. In the laboratory study 174 papers were inspected and bookmarked over a total of 625 bookmarks (27.8%), while in UMAP 2013 conference 31 papers were inspected and then bookmarked over a total 103 bookmaking actions (30%). The analysis shows that there is no significant difference (chi-square = 0.03, df=1), p=0.86, between the proportions of papers inspected in the Venn diagram and then bookmarked between study 2 controllable recommender interface and the UMAP 2013 controllable recommender interface.

7.1.2.4 Summary of Results Comparing Laboratory and UMAP 2013 User Behavior

Using statistical significance tests, three metrics were analyzed to compare the users' behavior between the realistic setting of UMAP 2013 and the laboratory setting of Study 2.

We observe that the realistic and the laboratory setting do not differ in terms of the proportion of people who use the sliders to filter papers, as well as in the proportion of papers that were inspected in the Venn diagram and then bookmarked. However, in the real setting, the proportion of papers bookmarked from the list produced after filtering with the sliders is considerably smaller compared to the talks bookmarked from the initial list of recommendations.

There are two possible explanations for this. The initial list of recommendations is good enough and then users do not need to update it to bookmark papers, or it can imply that the users do not perceive an improvement of the recommended items after controlling the list with sliders, so they explore this action but do not use it for bookmarking.

7.1.3 Results of Post-Conference survey

A post-conference survey was sent by e-mail to the users of the recommender interface in both HT and UMAP 2013. 4 attendees of the HT conference answered the survey and 8 attendees of the UMAP 2013 conference did it. Table 23 shows three columns: (1) the statements of the survey, (b) the average user agreement on the statements about the controllable interface in Study 2 (40 answers), and (c) the average user agreement on the statements about the controllable interface in both HT and UMAP 2013 (12 answers)

Table 23. Results of the post-conference survey, it also shows the results of the post-session survey in Study 2 (*=significant at p<0.05, ** p<0.01, using Mann-Whitney Test)

Statement	Controllable Interface in Study 2	HT and UMAP study
1. I understood why the talks were recommended to me.	4.05±0.09	3.58±0.38
2. The talks recommended matched my interests.	3.5±0.09	3.75±0.22
3. The talks recommended were diverse.	3.93±0.11	3.5±0.19
4. I became familiar with the recommender interface very quickly.	****4.45±0.09**	**3.67±0.28**
5. I lost track of time while I was using the recommender interface.	2.8±0.17	2.33±0.26

6. Overall, I am satisfied with the recommender interface.	**4.28±0.09	**3.67±0.22**
7. The recommender made me more confident that I didn’t miss relevant talks.	3.9±0.11	3.33±0.31
8. I would use this recommender system again for another conference in the future.	4.23±0.09	4.08±0.15
9. I would suggest my colleagues to use this recommender system when they attend a conference in the future.	4.28±0.09	3.92±0.23
10. I do not think that a social conference support system -like Conference Navigator- needs Talk Recommendation functionality.	1.93±0.17	1.5±0.23
Questions referring to the interface		
11. I felt in control of combining different recommendation methods by using the sliders.	4.03±0.13	4±0.21
12. The ability to control the recommendation methods increased my satisfaction with the list of recommended talks.	4.05±0.12	3.83±0.27
13. The ability to control the recommendation methods increases my trust in the list of recommended talks.	*3.95±0.11	**3.42±0.23**
14. When looking at the list of recommended talks I am interested to examine which recommendation method has been used.	4.03±0.15	3.42±0.31
15. I think the Venn diagram visualization helped me to understand why a talk was recommended.	4.08±0.13	4±0.28
16. I think the Venn diagram visualization was useful to identify talks recommended by a specific recommendation method or by a combination of recommendation methods.	*4.35±0.11	**3.58±0.34**
17. The ability to use the Venn diagram to examine the talks recommended increases my trust in the list of recommended talks.	3.83±0.12	3.5±0.26

The overall the user satisfaction with the recommender interface (statement 6) in a real setting is significantly smaller in the controlled laboratory study, p = 0.008. Three questions that show the largest gaps in users’ opinion between the laboratory and the real setting can help to understand this difference. First, there is an significant difference in becoming familiar with the interface (statement 4), p = 0.007. During the laboratory study, the users were given a short training about the capabilities of the recommender interface, while in the realistic setting users played freely without any kind of instruction. This might have prevented them from realizing the use of the Venn diagram for filtering, and it is reflected in a big gap between the positive perception of users in statement 16 compared to the less positive user perception in the HT and UMAP conferences, significantly different p = 0.019. This also might affect the perception for statement 1 "*I understood why the talks were recommended to me*" and the difference between the laboratory and the realistic setting. Finally, a user gave the following comment regarding the interface that supports the idea of not finding out all the capabilities of the Venn diagram, since this user does not have any log of clicking on the ellipses of the Venn diagram. The filtering feature could have helped him with his issue:

"While I was pleased with the recommendations, I was struggling a little bit with the mapping of the talks/papers in the diagram and their counterparts in the list."

7.2 SUMMARY AND DISCUSSION

This case study was aimed to explore how would users perceive and interact in a real setting with the recommender interface implemented for this dissertation without the support given in the laboratory study. Although the Study 1 was also conducted in the realistic setting of CSCW 2013, in this conference the preference elicitation mechanism was rather unnatural, users' bookmarks were not used to update the recommended lists, and most importantly, the filtering feature of the Venn diagram was not fully developed until Study 2. Three analyses were conducted and their results were presented in this chapter. First, the usage distribution analysis showed some differences between the laboratory setting and conference setting, the most important was the lack of use of the Venn diagram as an interactive filter. We associate this behavior to the lack of user training over visual widgets that are not widespread on web pages. Second, four metrics were tested to compare the proportional usage of the recommender interfaces in HT and UMAP conferences versus the laboratory setting. The proportion of people using the visual widgets of the recommender interface was not significantly different, but the proportion of papers bookmarked after using the sliders was significantly smaller than in the laboratory setting. Third and final, the results of a post-conference survey were presented and compared to the results obtained in the post-session survey of Study 2. The survey results revealed a bigger user perception of difficulty on becoming familiar with the interface, and a smaller overall and satisfaction. We think that the lack of user training might have affected significantly the user perception of the system, particularly how users' perceived the Venn diagram, ignoring their filtering capabilities (with the *clickEllipse* action) in addition to inspectability (with the *hoverCircle* action). This result is aligned with previous literature that has shown the challenge of engaging users to new visual features compared to familiar representations when introducing navigational tools in collections of documents. Chaopanon (2001) introduced a novel document navigational tool and the lack of significant usage compared to a traditional navigation interface is explained in big part by cognitive strain (Albers, 1997). According to Albers (1997), learning new methods and options requires additional work and remembering, and users tend to work to optimize their cognitive resources rather than maximizing their work output.

8.0 SUMMARY AND CONCLUSIONS

This dissertation has investigated the effects of user controllability in a hybrid recommender implemented on a conference support system. The purpose of this dissertation is to advance the investigation of human factors in recommender systems, an area that had been partially overlooked until the last four years, eclipsed by the fast progress of algorithms that improve off-line prediction accuracy but do not necessarily improve the user experience (McNee et al., 2006a). On studying the factors that influence the user experience, this investigation addresses four research question that can be summarized in two broad aspects:

- The role of user controllability on user engagement and overall user experience in a hybrid recommender system.
- The impact of user characteristics over the role of controllability on user engagement and experience in a hybrid recommender system.

The aforementioned aspects were addressed by conducting two user studies that contrasted the use of a controllable with a non-controllable recommender. In addition, an exploratory analysis describes the use of the controllable recommender interface in two conferences, in order to discover implementation details that may hinder the generalizability of results obtained in a laboratory setting. Each user study had a particular focus and helped to answer different research questions. The evaluations were addressed with usage metrics (actions and time), objective metrics (accuracy and rank metrics), and subjective measures (surveys). In addition, exploratory behavioral analysis was conducted to enhance the understanding of participants ‘experience with the interface. This multimode evaluation is one contribution of this research, since many previous studies focus on only one or two of the dimensions of evaluation described.

Although addressing different research questions, all the studies were conducted in Conference Navigator, an online web system that supports conference attendees. The studies have in common the utilization of a novel controllable recommender interface that is introduced as an important outcome of this research. The controllable recommender interface provides a visual interactive widget inspired in Venn diagrams that allow the users to fuse, inspect and filter recommendations. This is complemented by another visual widget composed of sliders that allow the user to control the fusion of recommender methods.

8.1 CONTRIBUTIONS AND IMPLICATIONS

8.1.1 Impact of the Visual Interface on User Engagement

The first research question of this dissertation is about the effect of the controllable recommender interface on user engagement. Measuring user engagement in online systems is a difficult endeavor, and relying only on behavioral metrics, such as clicks or time spent, is difficult to interpret. Hence, the use of diverse and different metrics is strongly suggested (Attfield et al., 2011). For this reason, in every study we relied on behavioral and subjective metrics to assess engagement.

In the first field study, users demonstrated interest in using the controllable interface by bookmarking talks, actively rating and exploring the visual components available, and also returning to the controllable interface. Although the users were more active with the controllable interface than with the non-controllable one (for instance, 31.81% came back to the controllable interface vs. 13.33% in the baseline), the effect was not big enough and only a few of these metrics were significantly larger than the non-controllable condition (talks bookmarked and number of clicks). Regarding subjective metrics, significant evidence was found. Users were more likely to agree that a recommender system was needed when using the controllable interface than when using the non-controllable one.

More evidence was found in the laboratory Study 2. Although the use of behavioral actions to measure engagement is limited if compared to a field study (we cannot count how many times the user comes back), we found an effect of the treatment (the controllable interface) on the amount of time spent on the task. When the controllable interface was presented as the second one (after the non-controllable interface), the users spent significantly more time than the opposite case, when participants used the non-controllable interface after the controllable one. We identify two possible explanations: users were significantly less efficient in the task because it required them more effort to finish, or possibly users were more engaged, so that they spent more time exploring the interface while finishing the task. After combining these measures with the results of the survey, we have more evidence of the second hypothesis. In the post-study survey, only 17.5% perceived that the controllable interface required more effort than the non-controllable one. In addition, when the controllable interface was presented second, users had a significantly larger agreement on the statements related to “use the interface again” and “recommending it to colleagues”.

These results indicate that there is an effect on user engagement in the interface, but the fact that the effect was stronger when users tried the non-controllable interface first has interesting implications. First, this might indicate the need to become familiar with the basic features of the interface before engaging with it. From a different perspective, it might also indicate that the visual features become more helpful after the user has become tired by performing a task of identifying relevant items without any support beyond the content.

In terms of measurement, it would be interesting to further evaluate the user engagement using recently investigated metrics (Dupret & Lalmas, 2013) such as absent time (in a real setting), time to first click, and click position.

8.1.2 Impact of the Visual Interface on Performance and User Experience

Results on the evaluation of Study 2 support the positive effects of the controllable interface compared to the non-controllable one. Although the interface did not have a direct effect on the usage metrics *talks explored* and *time spent*; users in the controllable interface effectively improved their mean average precision (MAP) and the perception of understanding why the talks were recommended. Other effects were larger only when the controllable interface was presented after the non-controllable one, i.e., the order in which the interface was presented played an important role. In this case, the perception of finding the interface easy to learn, expected future use and overall satisfaction with the interface are significantly higher in the controllable interface. This result is interesting because might be an example of the effect of cognitive strain, as explained by Albers (1997) and eventually described by (Chaopanon, 2001). The effect predicts that experiments assessing “new” navigational tools will continue to be biased in favor of the tools with which the users are already familiar. According the Albers, learning new methods and options requires additional work and use of memory, and as summarized by Chaopanon, “users tend to work to optimize their cognitive resources rather than maximizing their work output.”

It is also important to note the result from comparing the average user rating on different fusions of recommendations. The result indicates that the controllable interface affected the relative perception of the relevance of different fusions of recommender methods within the same interface. Further analysis is required to understand which specific feature or group of features produced the aforementioned effect, but the implication is important because it supports recent results on the importance of visualization, interactivity, and user intervention in recommender systems.

Another important result originated in the qualitative analysis of users’ opinions at the end of Study 2. Apart from many subjects praising the Venn diagram, the single most important feedback was that the visual controls make the interface unnecessarily complex if the list of recommendation is not very long. This indicates in which scenarios and domains the controllable features presented by sliders and by the Venn diagram would be most beneficial for information filtering. In order to answer this question, research can compare the users’ perception of the controllable interface, given either a short or long list of recommendations. This can be part of the inherent characteristics of different domains, so it might be possible that the controllable recommender proposed in this study has a stronger effect in a domain such as movie or music recommendations (with thousands or millions of possible items) rather than a small-to-medium size conference, on the order of a hundred papers to recommend.

8.1.3 Role of User Characteristics

The influence of user characteristics on several metrics supports previous findings but also reveals additional effects that should be further investigated to probe their generalizability. The effect of trusting propensity and experience in the domain were found relevant in this study and they have already been found relevant in shaping user behavior in previous studies. In our study, trusting propensity was found to increase the positive perception of the usefulness of the Venn diagram and the expertise on the domain increased the perception of diversity and decreased the feeling of being immersed in the task, i.e., decreased the user engagement with the system.

Being a native speaker had an influence by decreasing the number of actions on the sliders and on the Venn diagram, though it also increased the perception of understanding why the talks were recommended.

As a product of an exploratory factor analysis, we separated trusting propensity in general from the trusting propensity in recommender systems. The second one actually measures whether users had previously followed suggestions from traditional recommender services. In conjunction with previous usage of Conference Navigator, they increased the number of actions on the slider widget. Subsequently, the increased usage of the sliders widget decreased the perception of the usefulness of the Venn diagram, unlike increased usage of the Venn diagram, which enhanced that perception. This result highlights the competition between the sliders and the Venn diagram in explaining the recommended items, although both widgets were designed to be complementary.

Finally, an interesting result is the role of gender. In the final comments males expressed feeling less likely to perceive the Venn diagram as a good feature to filter and inspect recommendations; most of them preferred the sliders. The post-session survey shows that males did not find the Venn diagram useful for understanding why a talk was recommended whereas females did. Although further studies should be conducted to generalize this result, it would be interesting to check whether gender can influence the perception of certain visual representations that can further more influence the user experience in recommender systems.

8.1.4 Real Setting vs. Controlled Laboratory Study: The Importance of Training

The exploratory behavioral analysis conducted in the realistic setting of two conferences (HT and UMAP 2013) confirmed some expectations and revealed differences between interactions in a controlled laboratory setting and a realistic scenario. First, users actually engaged with the system in a realistic setting without training and without extensive promotion, with the exception of an e-mail to conference attendees and ads in the homepage of Conference Navigator. Users explored the interface by filtering with the sliders and hovering over the Venn diagram to inspect the recommended items. However, users dismissed the use of some actions available in the recommender interface: filtering the list by clicking on the Venn diagram areas and intersections.

We believe that the lack of user training in the controllable recommender interface affected user action sthat did not have a straight forward affordance (the filtering actions on the Venn diagram), despite being particularly praised by participants in Study 2. This result can be explained by the same effect experienced by Chaopanon (2001) when introducing a new "navigational information" widget: cognitive strain. Users were more likely to use familiar tools that maximize their cognitive resources rather than the task output. The implication here is to consider a significant user training period for getting familiar with the interface before introducing visual or interactive features that depart from traditional representations.

8.2 FUTURE WORK

8.2.1 IR Metrics on Traditional versus Interactive Environments

In this dissertation, IR metrics were used to compare the user experience between a static and an interactive list of recommendations. Although one of the metrics, mean average precision (MAP), was significantly higher in the interactive interface, a conclusion of better user experience based solely on this result is far from definitive. The conclusions of this dissertation with respect to the effect of the controllable recommendations were supported with several additional measures, but if one considers only IR metrics such as MRR, MAP or nDCG, the issue is that they have been created for static lists of recommendations and not for interactive sessions. One way to extend the work of this dissertation is analyzing the accuracy and ranking results using recently created metrics such as sDCG (Järvelin et al., 2008) that consider the interactive nature of a session rather than assuming a static list of recommendations.

8.2.2 HCI research on Recommender Systems

Several studies in the last 5 years (Bostandjiev et al., 2012; Gretarsson et al., 2010; Knijnenburg, Bostandjiev, et al., 2012; O'Donovan et al., 2008; Pu et al., 2011) have highlighted the importance of studying the HCI dimension in the research on recommender systems, which traditionally has focused on ways to improve the user experience by optimizing system characteristics such as prediction accuracy or item diversity via algorithms. This dissertation has shown how different visual representations and interactions (the use of sliders and an interactive Venn diagram for controllability) affect the user experience in a recommender system. The adoption of HTML5 by the most important browsers and the availability of several Javascript libraries that facilitate the implementation of rich visualization–such as D3.js, Raphael.js, or InfoVis-open a large spectrum of possible visualizations and user interactions that can be studied in different applications and domains in the present and future. Therefore, an interesting area of future work is studying the effect of different visual representations (graphs, sunbursts, chord diagrams) and interface interactions on the user experience with a recommender. Previous research on visual representations and their influence on decision making (Lurie & Mason, 2007) can be used to define appropriate research questions applied to this area.

Moreover, the gap between exploratory search (White & Roth, 2009) and interactive recommender systems is rather narrow, and combining analysis techniques frequently used in one area that have not been used in the other can benefit both areas of research. For instance, Hidden Markov Models have been used to study exploratory search (Yue, Han, Jiang, & He, 2012) and this technique could be used to extend the behavioral analysis of an interactive recommender system like the one presented in this dissertation.

Another important topic for future research is to enhance the study of user interactions with the interface for mobile devices. In a mobile device there are frequently many potential user actions that can help to identify user engagement: zooming with fingers, using 1 or 2 fingers to slide lists or specific elements, sharing content through different online social services (Guo, Jin, Lagun, Yuan, & Agichtein, 2013). These set actions increase the set of signals used to assess user preference compared to a traditional device, enriching the user feedback utilized to improve recommendation methods.

APPENDIX A

SURVEYS

A.1 PRE-SURVEY

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

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

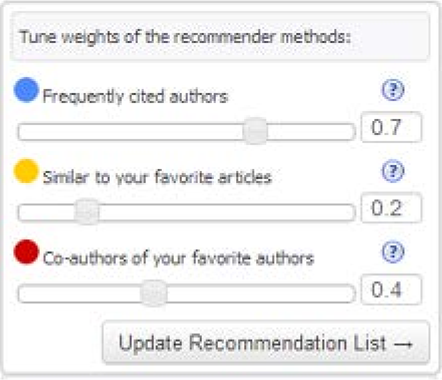
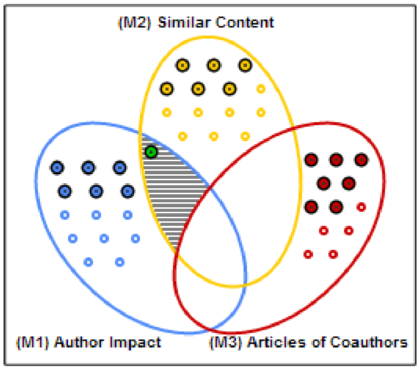
19. I know of one or more methods used to produce recommendations in a system

Strongly disagree / Disagree / Neutral / Agree / Strongly Agree

A.2 POST-SESSION SURVEY (CONTROLLABLE CONDITION)

This survey was used in all the studies described in this dissertation.

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="3">
Tune weights of the recommender methods:
● Frequently cited authors 0.7
● Similar to your favorite articles 0.2
● Co-authors of your favorite authors 0.4
Update Recommendation List →
Sliders</td>
<td>
Venn Diagram</td>
</tr>
<tr>
<td></td>
<td colspan="3">&lt;&lt; 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>
</table>

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
	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.

A.3 POST-SESSION SURVEY (NON-CONTROLLABLE CONDITION)

This survey was used in all the studies described in this dissertation

<< 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.	

A.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]

APPENDIX B

CODE FOR STATISTICAL TESTS

B.1 REGRESSION ON NUMBER TALKS EXPLORED

```
library(glmmADMB)

# =========== Comparable actions among interfaces

glm_exp3_std   <-   glmmadmb(comp_nbr_actions   ~   conditionID   +   order   +
      conditionID:order + program + gender + native_speak + age + std.trust +
      std.trustRecs + std.recsysExp + std.researchExpDom +  af.binCNUse + ( 1
      |af.userID) ,data=userec.df3,family="nbinom",zeroInflation=FALSE);

summary(glm_exp3_std)
```

B.2 REGRESSION ON TIME SPENT PER TASK

```
library(glmmADMB)

# ========== Time Spent

glm_time_std <- glmmadmb(time_spent ~ conditionID + order + conditionID:order
+ program + gender + native_speak + age + std.trust + std.trustRecs +
std.recsysExp + std.researchExpDom +  af.binCNUse + (1|af.userID)
,data=userec.df3,family="Gamma",zeroInflation=FALSE);
summary(glm_time_std);
```

B.3 REGRESSION ON MEAN AVERAGE PRECISION

```
library(lme4)
library(lmerTest)

# ========== MAP

lm_map <- lmer(map  ~ conditionID + order + conditionID:order + program +
gender + native_speak + age + std.trust + std.trustRecs + std.recsysExp +
std.researchExpDom +  af.binCNUse + (1|af.userID) ,data=userec.df9, REML =
FALSE);

summary(lm_map)
```

B.4 REGRESSION ON SURVEY ITEMS

```
library(lme4)
library(lmerTest)

# ------------------------- Subjective metrics -------------
lm_UND <- lmer(UNDERSTOOD  ~ conditionID + order + conditionID:order +
program + gender + native_speak + age + std.trust + std.trustRecs +
std.recsysExp + std.researchExpDom +  af.binCNUse + (1|af.userID)
,data=userec.df3);

summary(lm_UND);

lm_REL <- glmer(RELEVANT  ~  conditionID + order + conditionID:order +
program + gender + native_speak + age + std.trust + std.trustRecs +
std.recsysExp + std.researchExpDom +  af.binCNUse + (1|af.userID)
,data=userec.df3);

summary(lm_REL);

lm_DIV <- glmer(DIVERSE  ~  conditionID + order + conditionID:order + program
+ gender + native_speak + age + std.trust + std.trustRecs + std.recsysExp +
std.researchExpDom +  af.binCNUse + (1|af.userID) ,data=userec.df3);

summary(lm_DIV);
```

B.5 CODE FOR ONE SAMPLE TEST OF PROPORTIONS

```
prop.test(173,616,p=0.14)

        1-sample proportions test with continuity correction

data:  173 out of 616, null probability 0.14
X-squared = 100.3256, df = 1, p-value < 2.2e-16
alternative hypothesis: true p is not equal to 0.14
95 percent confidence interval:
 0.2460211 0.3184455
sample estimates:
        p
0.2808442
```

B.6 CODE FOR FACTOR ANALYSIS

```
# ==========================================================================
# FINAL USED IN THE DISSERTATION WRITING
# ==========================================================================
# http://cran.r-project.org/web/packages/polycor/polycor.pdf
library(MASS)
library(GPArotation)
library(mvtnorm)
library("psych")
corrmat.pre <- mixed.cor(p = presurvey[, c(6,8,15:20)],  polycor = TRUE)
presurvey.corr.pre <- as.matrix(corrmat.pre[['rho']][1:8,1:8])
f1.pre <- fa(presurvey.corr.pre,4,n.obs=40,fm="mle", rotate="varimax")
fa.diagram(f1.pre)
f1.pre$loadings
factor.scores( presurvey[, c(6,8,15:20)],f1.pre)
# ==========================================================================
```

BIBLIOGRAPHY

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.

Ariely, D. (2000). Controlling the Information Flow: Effects on Consumers' Decision Making and Preferences. *Journal of Consumer Research, 27,* 233-248.

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.

Chaopanon, W. (2001). *Semantics, complexity and capability: the use of integrated navigational tools for information finding in hypertext document space.* University of Pittsburgh.

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.

Dupret, G., & Lalmas, M. (2013). *Absence time and user engagement: evaluating ranking functions*. Paper presented at the Proceedings of the sixth ACM international conference on Web search and data mining, Rome, Italy.

Ekstrand, M. D., Kannan, P., Stemper, J. A., Butler, J. T., Konstan, J. A., & Riedl, J. T. (2010). *Automatically building research reading lists.* Paper presented at the Proceedings of the fourth ACM conference on Recommender systems, New York, NY, USA.

Farzan, R., & Brusilovsky, P. (2008). *Where did the researchers go?: supporting social navigation at a large academic*. Paper presented at the Proceedings of the nineteenth ACM conference on Hypertext and hypermedia, Pittsburgh, PA, USA.

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

Gou, L., You, F., Guo, J., Wu, L., & Zhang, X. (2011). *SFViz: interest-based friends exploration and recommendation in social networks*. Paper presented at the Proceedings of the 2011 Visual Information Communication - International Symposium, New York, NY, USA.

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.

Guo, Q., Jin, H., Lagun, D., Yuan, S., & Agichtein, E. (2013). *Mining touch interaction data on mobile devices to predict web search result relevance*. Paper presented at the Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval, Dublin, Ireland.

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.

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., Rao, N., & Kobsa, A. (2012). Experimental Materials Used in the Study on Inspectability and Control in Social Recommender Systems: UCI.

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.

Lamere, P. (2012). *I've got 10 million songs in my pocket: now what?* Paper presented at the RecSys.

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., Albert, I., Cosley, D., Gopalkrishnan, P., Lam, S. K., Rashid, A. M., . . . Riedl, J. (2002). *On the recommending of citations for research papers.* Paper presented at the Proceedings of the 2002 ACM conference on Computer supported cooperative work, New York, NY, USA.

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.

Minkov, E., Charrow, B., Ledlie, J., Teller, S., & Jaakkola, T. (2010). *Collaborative future event recommendation.* Paper presented at the Proceedings of the 19th ACM International Conference on Information and Knowledge Management, 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. (2009). *Collaborative filtering for social tagging systems: an experiment with CiteULike.* Paper presented at the Proceedings of the third ACM conference on Recommender systems, New York, NY, USA.

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.

Pazzani, M. J., & Billsus, D. (2007). Content-based recommendation systems. In B. Peter, K. Alfred & N. Wolfgang (Eds.), *The adaptive web* (pp. 325-341): Springer-Verlag.

Prajapati, B., Dunne, M., & Armstrong, R. (2010). Sample size estimation and statistical power analyses. *Optometry Today, 16*(07).

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.

Sahebi, S., Wongchokprasitti, C., & Brusilovsky, P. (2010). *Recommending research colloquia: a study of several sources for user profiling.* Paper presented at the Proceedings of the 1st International Workshop on Information Heterogeneity and Fusion in Recommender Systems, 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.

Troussov, A., Parra, D., & Brusilovsky, P. (2009). *Spreading Activation Approach to Tag-aware Recommenders: Modeling Similarity on Multidimensional Networks.* Paper presented at the Proceedings of the Workshop on Recommender Systems and the Social Web.

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.

White, R. W., & Roth, R. A. (2009). Exploratory search: Beyond the query-response paradigm. *Synthesis Lectures on Information Concepts, Retrieval, and Services, 1*(1), 1-98.

Yue, Z., Han, S., Jiang, J., & He, D. (2012). *Search tactics as means of examining search processes in collaborative exploratory web search*. Paper presented at the Proceedings of the 5th Ph.D. workshop on Information and knowledge, Maui, Hawaii, 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.