CONTROLLABILITY AND EXPLAINABILITY IN A HYBRID SOCIAL RECOMMENDER SYSTEM

by

Chun-Hua Tsai

B.B.A. in Information Management, Tamkang Univ., Taiwan, 2007

M.S. in MIS, National ChengChi University, Taiwan, 2009

Submitted to the Graduate Faculty of

the School of Computing and Information in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2019

UNIVERSITY OF PITTSBURGH

SCHOOL OF COMPUTING AND INFORMATION

This dissertation was presented

by

Chun-Hua Tsai

It was defended on

Auguest 22nd 2019

and approved by

Dr. Peter Brusilovsky, Professor

School of Computing and Information, University of Pittsburgh

Dr. Yu-Ru Lin, Associate Professor

School of Computing and Information, University of Pittsburgh

Dr. Konstantinos Pelechrinis, Associate Professor

School of Computing and Information, University of Pittsburgh

Dr. John O'Donovan, Associate Research Scientist

Department of Computer Science, University of California, Santa Barbara

Dissertation Director: Dr. Peter Brusilovsky, Professor

School of Computing and Information, University of Pittsburgh

CONTROLLABILITY AND EXPLAINABILITY IN A HYBRID SOCIAL RECOMMENDER SYSTEM

Chun-Hua Tsai, PhD

University of Pittsburgh, 2019

The growth in artificial intelligence (AI) technology has advanced many human-facing applications. The recommender system is one of the promising sub-domain of AI-driven application, which aims to predict items or ratings based on user preferences. These systems were empowered by large-scale data and automated inference methods that bring useful but "puzzling" suggestions to the users. That is, the output is usually unpredictable and opaque, which may demonstrate user perceptions of the system that can be confusing, frustrating or even dangerous in many life-changing scenarios. Adding controllability and explainability are two promising approaches to improve human interaction with AI. However, the varying capability of AI-driven applications makes the conventional design principles are less useful. It brings tremendous opportunities as well as challenges for the user interface and interaction design, which has been discussed in the human-computer interaction (HCI) community for over two decades. The goal of this dissertation is to build a framework for AI-driven applications that enables people to interact effectively with the system as well as be able to interpret the output from the system. Specifically, this dissertation presents the exploration of how to bring controllability and explainability to a hybrid social recommender system, included several attempts in designing user-controllable and explainable interfaces that allow the users to fuse multi-dimensional relevance and request explanations of the received recommendations. The works contribute to the HCI fields by providing design implications of enhancing human-AI interaction and gaining transparency of AI-driven applications.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

PREFACE

This work would not have been possible without supports from many academic professionals. I would like to express my deepest appreciation to my academic advisor, Dr. Peter Brusilovsky, for his guidance through each stage of the process. I am very lucky to have him in my journey of pursuing my doctoral degree. I would also like to extend my deepest gratitude to my dissertation committee, Dr. Yu-Ru Lin, Dr. Konstantinos Pelechrinis and Dr. John O'Donovan, for all their help and feedback on my dissertation works.

I very much appreciate all the professors who even offer me any kind of help, in particular, thank goes to Dr. Dmitriy Babichenko, Dr. Shih-Yi Chien, Dr. William W. Clark, Dr. I-Ling Cheng, Dr. Rosta Farzan, Dr. Daqing He, Dr. I-Han Hsiao, Dr. Hassan Karimi, Dr. Yihuang Kang, Dr. Prashant Krishnamurthy, Dr. Pei-Ju Lee, Dr. Yiling Lin, Dr. Shaghayegh Sahebi, Dr. Martin B.H. Weiss, and Dr. Vladimir Zadorozhny (in alphabetical order). Thank you for your support and guidance. Special thanks to Dr. John Carroll and Dr. Mary Beth Rosson, who grants me an opportunity to advance my academic career.

I am extremely grateful to have many warm and supportive staff and colleagues. Many thanks to Ms. Kelly Shaffer and Dr. Sung-min Kim who are my best friends and are always there for me. My thank goes to all my colleagues, included Mr. Kamil Akhuseyinoglu, Ms. Nuray Baltaci Akhuseyinoglu, Mr. Yongsu An, Mr. Jordan Barria-Pineda, Mr. Kim Chau, Ms. Yu Chi, Dr. Jidapa Kraisangka, Mr. Xin Liu, Dr. Di Lu, Mr. Rui Meng, Ms. Khushboo Thaker, Dr. Chirayu Wongchokprasitti, Dr. Xidao Wen, Ms. Fanghui Xiao, Ms. Fan Yang, Ms. Danchen Zhang, Mr. Sanqiang Zhao, and Ms. Ning Zou. There many more friends at Carnegie Mellon University and the community where I belong. Special thanks to Dr. Huan-Kai (Pumbaa) Peng who provided useful feedback to my preliminary exam. A million thanks to all of you. I wish you all the success today and always!

I would like to acknowledge the help of all the research participants in my experiments, all the reviewers of my publications, and all the people who ever interact with me in conferences or other events. I will not be able to complete this dissertation without your effort and help. I gratefully acknowledge the assistance from the Taiwan government, which sponsored me the *"Government Fellowship for Studying Abroad"*. I am proud of being a citizen of Taiwan.

Finally, I would like to extend my deepest gratitude to my wife, Ms. Chuntzu Sage Liu, who is my best friend and soulmate, stay with me in countless nights of self-doubt, fear, and uncertainty. It is hard to image to walk through this journey without you. Many thanks to her to have us a beautiful son, Benjamin Tsai. All my achievements become meaningful because of you. Love you and we will continue to support each other and grow from each other in the future. Thanks should also go to my parents, sister, brother. All my friends in Taiwan and around the world. I want to share my happiness with you.

P.s. It is worth to mention Puzzle and Bak Kut Teh, they are my lovely cats. :-)

Chun-Hua Tsai
State College, PA, USA, 2019

1.0 INTRODUCTION

1.1 MOTIVATION

The growth in artificial intelligence (AI) technology has advanced many human-facing applications. *Recommender system* is of the promising sub-domain of AI-driven systems, which has been applied to many different real-world applications. For example, *social recommendation* is one of the popular use cases, which has been widely adopted in many e-commerce platforms or social media. Providing social recommendation is aims to filter out "irrelevant" information so the users can reduce the efforts of decision making, e.g., purchase an item online or follow a new friend on social media. The AI-driven social recommender systems are usually empowered by large-scale data from multiple data sources and automated inference methods that bring effectively but "puzzling" output to the users [4]. The output is usually unpredictable and opaque, which may demonstrate user behavior that can be confusing, frustrating or even dangerous in many life-changing scenarios. This dissertation presents my exploration of the value of bringing controllability and explainability to a hybrid social recommender system.

Hybrid social recommender systems attempt to improve the quality of recommendations by engaging with several recommendation sources or approaches [12]. While hybrid recommenders are known for their performance, their sophisticated computational processes are less transparent than those of other approaches to hybridization. Even the simplest parallel hybridization approach, which fuses together multiple recommendation sources with different weights might opaque to the users [12]. Usually, the optimal fusion weights are trained or learned using ground truth data to optimize for the best overall performance. The result might be optimal, but not easy to comprehend. As a result, user actions affecting one of

the sources might result in confusing changes to the final recommendation list. Moreover, this hidden optimal fusion might not even work well for the users. In many real-world scenarios, a hybrid social recommender system can be preferred for different user needs [132] or by multi-stakeholder [1]. For example, in an event-based social recommender, the event attendees may look for other people for a range of reasons, i.e., a conference attendee may wish to re-connect with an acquaintance, find new friends with similar research interests, or just find someone with whom to share a ride to the airport. This diversity of information needs makes it difficult to generate a static ranked social recommendation list that fits all cases [17]. While an *optimal* static fusion could provide the best ranking across the cases, it might be *sub-optimal* in each specific case.

These challenges were recognized and addressed in the new generation of research involving *interactive recommender systems* [56]. These systems attempted to make the recommendation process more transparent by visualizing some aspects of the process and offered the user some form of control over the process. Starting with the pioneering work on PeerChooser [98], several attempts were made to produce more transparent and controllable recommender systems. In these systems, users were allowed to "influence" the presented recommendations by interacting with different types of visual interfaces. Several studies of interactive recommender systems demonstrated that users appreciate *controllability* in their interactions with the recommender system [70, 55, 104, 29]. It has also been shown that the visualization has helped users to understand how their actions can impact the system [62], which contributes to the overall *inspectability* [70] or *transparency* of the recommendation process [126].

A social recommender system is generating recommendations with user-generated data and algorithms. The "reasons" of the receiving the recommendations usually stay in a *black box* [58] that the user has little understanding about the mechanism behind the system. A lower transparency system has been proved association with user satisfaction negatively [126]. To gain the transparency, the study of [58, 126] suggested providing proper *explanations* to help the user to understand the reasoning process of the generated recommendations. The approach aims to provide more details that make the users realize the reasons for receiving the recommendations. In many user-centered evaluations [58, 70, 125], the explanation

positively contributes to the user experience, i.e., trust, understandability, and satisfactions [41, 126]. However, ongoing advance in AI techniques generates tremendous opportunities as well as challenges for the user interface and interaction design, which has been discussed and debated in the human-computer interaction (HCI) [4]. For instance, how can we design an effective explanation interface? How can the explanations affect the recommender system controllability? How does the user adopt these explanations in decision-making and information seeking processes?

Enhancing explainability in recommender systems has drawn more and more attention in the field of Human-Computer Interaction (HCI). Further, the newly initiated European Union's General Data Protection Regulation (GDPR) required the owner of any data-driven application to maintain a "right to the explanation" of algorithmic decisions [32], which urged to gain *transparency* in all existing intelligent systems. Self-explainable recommender systems have been proved to gain user perception on system transparency [125], trust [107] and accepting the system suggestions [70]. The problems of controllability and transparency of hybrid recommendation processes have been explored in several projects [70, 104, 29], however, these projects focused on either transparency or controllability of the *fusion component*. My own experience demonstrated that visual interfaces for user-controlled fusion cannot assure that these users will understand the underlying rationale of each contributing data or methods; namely, how the recommendation has been made [62]. If a recommendation mechanism is too complicated for non-professional users to comprehend, considerable transparency could be achieved by *explainability*, i.e., the system may *justify* why the recommendation was presented [132, 29]. I believe that to increase the transparency of social recommender systems; interactive user interfaces should be augmented with multiple kinds of explanations for each recommendation source or engine. Several interfaces and approaches to provide explanations have been proposed and studied to assess the improvement of user satisfaction and other system aspects [58, 125]. Explaining recommendations can achieve different *explanatory goals* through single-style or hybrid explanations [75].

However, little is known about how the user will interact with the system when both the fusion process and reasoning process are transparent. In this dissertation, I investigated the effects of adding user control and visual explanations to an interactive hybrid social

recommender system. I proposed *Relevance Tuner+* to provide a controllable and explainable interface for the user to fuse social recommendations from multiple sources, using the *Conference Navigator* platform [10]. My contribution covers several aspects of transparency of recommendation. First, I proposed and evaluated novel user-controllable intelligent user interface and explanatory visualizations to enhance the controllability and explainability of a social recommender system. Second, I discussed the prospect trade-off between the transparent fusion and reasoning processes, which implied the interaction effects between controllability and explainability in a social recommender system. Third, I present implications for the design of the user interface to enhance user controllability and explainability based on these results. My work has great potential to extend to other recommender systems beyond this context, thereby making significant contributions to the research topics on the intelligent user interface (IUI), fair, accountable, and transparent (FAT) recommender systems (RS) as well as the domain of Explainable AI (XAI).

1.2 RESEARCH QUESTIONS

The integration of controllability and explainability in a hybrid social recommender system allowed me to address two important research questions (RQs).

- **[RQ1] How do controllability and explainability affect the *user perception*, *user experience* and *user engagement* with a hybrid social recommender system?**

The main research question in this dissertation is to explore how do controllability and explainability affect the user perception, user experience and user engagement of the social recommender system. In this dissertation, I bridge this gap by building user interfaces that provide social recommendations in ways that are compatible with effective user interaction and that can transparently present the recommendation reasoning process. I start by building a set of controllable user interfaces that supports the users to effectively fuse the multi-dimension recommendation relevance for different social information needs. I then

follow a stage-based participatory process to design a set of explainable user interfaces by exploring the recommendation reasoning process, user preference, and user performance. To assure the effectiveness of the interfaces, I conduct several user studies [132, 134, 136, 135]. These studies focused on exploring an effective design of user-controllable and explainable interfaces, which allows users to fuse multiple relevances for influencing the presentation of recommendation as well as seeking the explanation of the received recommendation. Finally, I integrated the evaluated controllable and explainable user interfaces into a hybrid social recommender system. I discussed the effects on objective and subject metrics as well as examined the system though a user-centric evaluation framework.

- **[RQ2] Is there an interaction effect between controllability and explainability in a hybrid social recommender system?**

The secondary research question is to identify the interaction effect between controllability and explainability, which enhances a different level of system transparency. Many studies have been investigated the effects of building controllable [98, 104, 56, 34] or explainable user interfaces [58, 126, 9, 140, 104] for recommender systems. However, neither a direct comparison of these interfaces nor the interaction effects between controllability and explainability were studied, i.e., little is known about how the user will interact with the system when both the fusion process and reasoning process are transparent. In my experience [132, 134, 136, 135], I found offering controllability and explainability in a hybrid social recommender system can improve the user perception of control and trust, respectively. It shows preliminary evidence that the different levels of system transparency may contribute to different user perceptions as well as be adopted in different information-seeking tasks. Since controllability and explainability provide a different level of transparency to the system, it is crucial to understand how a user will interact with a system that provides multiple levels of transparency.

1.3 CONTRIBUTIONS

The contributions of this dissertation can be summarized in four-fold.

1. This dissertation systematically explores the effective design of the user-controllable and explainable interfaces and demonstrates the value of integration these interfaces into a hybrid social recommender system. The experiment results provide empirical evidence to explain the user experience and interaction pattern in the proposed controllable and explainable interface design. I build a conceptual framework to explain the user interaction patterns which can be contributed to the theoretical confirmation and extension. The design can be easily extended to other recommendation systems in different contexts.
2. This dissertation introduced several real-world information-seeking tasks in human experiments. The experiment data would be beneficial in explaining the social information seeking and exploration of applying social recommender in different information needs. It provides a data-driven analysis of how users leverage a recommender system when the recommendation fusion and reasoning is transparent. These findings provided design implications that can be extended in a different context of AI-driven applications.
3. This dissertation presents a pioneer work that explores the use case of bringing effective controllable and explainable interfaces in a hybrid social recommender system, which contributes to the transparent on both of the recommendation fusion and reasoning processes. This work helps to understand the effect of user perception on a different level of system transparency. The interaction effect between the controllable and explainable interfaces are also discussed in this dissertation.
4. This dissertation has great potential to extend to other recommender systems beyond this context, thereby making significant contributions to the research topics on the intelligent user interface (IUI), fair, accountable, and transparent (FAT) recommender systems (RS) as well as the domain of Explainable AI (XAI).

1.4 DISSERTATION ORGANIZATION

The chapters of this dissertation are organized as follows:

"**Chapter 2: Related Works**" presents a literature review on the related research topics and works of this dissertation.

"**Chapter 3: Research Platform and Experiment Settings**" presents introduction of an event-based social recommender system, the recommendation models and the experimental measurements.

"**Chapter 4: User Controllable Interfaces for a Hybrid Social Recommender System**" presents the pre-study results on exploring effective user-controllable interfaces for a social recommender system. The content of this chapter were partially published in [132] and [137].

"**Chapter 5: Designing Explanation Interfaces Using Stage-based Participatory Design Approach**" presents the investigation of designing explanation interfaces for a social recommender system. The content of this chapter were partially published in [133].

"**Chapter 6: Evaluating Prototypes of Explanation Interfaces**" presents the evaluation of the proposed explanation interfaces for a social recommender system. The content of this chapter were partially published in [136, 135].

"**Chapter 7: Explaining Social Recommendations in an Interactive Hybrid Social Recommender**" presents the experiment results of adding explainable user interfaces to an interactive hybrid social recommender system. The content of this chapter were partially published in [136].

"**Chapter 8: Evaluating Explanation Interfaces using Crowdsourcing Approach**" presents the evaluation of the proposed explanation interfaces using crowdsourcing approach.

"**Chapter 9: Controllability and Explainability in a Hybrid Social Recommender System**" presents a lab-controlled, large scale user study of the proposed user controllable and explainable interface.

"**Chapter 10: Conclusions**" presents the conclusion of this dissertation and the discussions of limitation and future works.

2.0 RELATED WORKS

2.1 SOCIAL RECOMMENDER SYSTEMS

Social media is a place to share various kinds of information as well as react to information shared, by using tags, likes, comments, or votes. The rapid growth of both content and data can be leveraged to provide accurate item recommendations, but it can also cause information overload, which makes it harder for users to filter interesting or relevant content [50]. This problem has been addressed in two different ways, forming two main streams of research on *social recommendation*. The first stream focuses on improving the traditional recommender approaches (i.e., ranking-based collaborative filtering [69]) by using various kinds of social data available in social media systems such as social links [82], tags [7], or reviews [99]. The second stream focuses on people recommendations, which aims to reduce social overload on making people-to-people connections. Within social media, recommendations of "people you may know" [52] or new and interesting people to follow [51] could remarkably improve access to relevant information. However, the value of people recommendation is not limited to improving information access. For example, an online dating service can enable strangers to establish new personal relationships, while academic recommendation helps to find co-authors and project collaborators [117]. In this dissertation, I follow the second stream (people recommendation) and explore it in a primarily academic context.

The people recommendation aims to reduce the social overload on people-to-people connections. For example, the social matching application to facilitate the engagement between people, like online dating service or *people you may know* in social media. The connections can be categorized as four relations [50]:

- *Symmetric v.s. Asymmetric*: The symmetric relationship requires the agreement between the two parties in the connections. For example, when one user A sends an invitation to the other user B, it requires approval from user B to establish the connections. The friend request on Facebook is considered an asymmetric relationship. On the other hand, the connection could be asymmetric if the relationship is one-way. For example, when user A sends an invitation to the user B, the single connection is established without the approval from the other party. The following request on Twitter can be considered as an asymmetric relationship.
- *Confirmed v.s. Non-Confirmed*: The symmetric or asymmetric connections can be with or without confirmation from both parties. Typically, the symmetric relationship requires confirmation from both parties. An asymmetric may (not) require a confirmation from one side.
- *Ad-Hoc vs. Permanent*: The connections can be established for permanent or just for particular time or events. For example, on Facebook, once you agree on the friend request, then you become friends until one side decides to terminate the connections. In some cases, we may decide to establish the connections for particular events, e.g., the connections between the participant of a summer hiking trip.
- *Different Domains*: The domains decide the property of the social recommender system. For example, in Facebook, the connection would be friends, colleagues, or families. In social media like Linkedin, the connection belongs to a more professional relationship.

Users could seek social recommendations based on different information needs. For example, a user may approach the social recommender system for finding new friends with similar interests or re-connect the acquaintance for chatting or browser. The scenario difference would change the method when recommending social connections. The study from [17] presented the social coverage in different recommendation algorithms, i.e., the content-based method would generate a list of social recommendations with more unknown users versus a social-based method would contain more known users in the list. An effective social recommendation is scenario-dependent, i.e., there may not be a "one-fit-for-all" approach for all social information needs. In this section, I will introduce three social recommendation examples based on the literature review.

First, *recommending people to connect with* is a well-known function of social media [35]. It establishes a mutual connection between users on social media, which builds the social network or social capital in cyberspace. The property of connecting is varied in many different domains. The social recommender systems play a role in providing suggestions for users to establish *mutual* connections, which require a mutual agreement from both parties. The mutual connections vary in different systems, e.g., it may be a friendship in a social networking service, professional networking between the employer and employee in an employment-oriented social networking service, or a co-authorship in a knowledge-based social network service. The connections can represent online as well as hyper-local relationship. For example, the event-based social system presented in [10], provided a "make a connection" function for connecting conference attendees and conference paper authors.

Second, since not all connections require mutual agreement, an important kind of recommendation is *recommending people to follow*. The goal is to establish a follower-followee relationship between the users. More specifically, the goal is to establish the follower and followee relationship between the users. For instance, on Twitter, the user can follow a friend, politician, or movie star, which is a one-way relation. The study of [48] attempted using a graph-based method to generate the recommendation on Twitter. The algorithm is applied random-walk techniques based on the network follower and followee. An offline experiment and online A/B testing were scheduled to test the approach, but how does the method can facilitate the user to follow the interested person is not reported in the paper.

Third, social media is not limited to affirm or enforce the existing social connections but may aim to extend the social network beyond the cycle. The third stream of social recommendation research focuses on *recommending strangers* to the user. It is a useful function in many different applications, e.g., seek advice from an expert, to find new co-work opportunities or to know new friends. It may be a challenging task due to the fact of cold-start issue. An efficient recommender system requires robust signals (user profile data) to predict the valuables item to the user. However, in a scenario for recommending strangers, the system may not have sufficient information to generate such suggestions. In the study of [53], the authors introduced a social networking service called *StrangerRS* that attempted to recommend the unknown user in a corporation. Their approach is to "present" the related

information between the two parties who may be potentially interested in connecting. A user interface was designed and deployed in a controlled field study. The experiment results showed the interface could increase the number of unknown people recommendation to users. However, how to apply the interface to improve the preference elicitation is a remaining open question.

2.2 RECOMMENDER SYSTEM INTERFACES

There are studies that tried to enhance the recommendation diversity through exploratory search interfaces. For instance, Orso et al. [100] introduced an interface for interactive search, which uses overlaying graphs representing different information sources. The studies of [67, 68] proposed interfaces which can present multi-faceted information on map or touch screen to diversify and explore the generated recommendations. The studies of [147] adopted dimensional-reduction techniques to project the multidimensional data in two or three dimensions for the purposes of visualization. SciNet interface [111, 42] recommends keywords spatially in an interactive visualization to help diversification in exploratory search in a visualized surface of polar coordinates. In the field of information retrieval, a similar idea has been discussed to give users the capability to navigating and visualizing the search results in a multidimensional image renderer [28]. The work of [79] further adopted a 3D item space visualization for presenting multi-faceted data and user preferences in a movie recommender using the collaborative filtering approach. The users can interactively manipulate the recommended item by adjusting the "landscape" on the map.

Providing a visual interface is another approach to solving the diversity recommendation problem. A visual discovery interface to increase the CTR rate in the e-commerce website [122]. An interface to use a two-column format to present the two side opinions of controversial subjects [83], which may reduce the distance of latent space among users' ideology [44]. Explainable interfaces can be used to justify the reason for recommendations for the user to actively access different contents [151, 150]. Pu et al. [107] designed an explanation interface to justify the recommendation result. The explanation is useful for the user to understand

the reason for getting a positive recommendation. The user can determine to explore more based on the explanation. Schafer et al. [113] proposed a user-controllable interface for the user to interactively change the ranking or feature weighting, for a better-personalized ranking. It is a common approach to adopt user interfaces to show the various recommendation result [36, 147, 60, 142]. In this section, I will introduce two examples that inspired my dissertation works.

First, Verbert et al. [140] proposed *TalkExplorer*, an interactive visualization developed on top of the Conference Navigator system (CN3) [10], which visualizing recommendation, tags, and users with the content-based and tag-based recommender. The goal of the system is to provide a talk recommendation of a conference with the explanation of the recommended item in a transparent way and to support exploration and control by the end-users. The target users are the conference attendees who need to explore the interesting talks to attend or find some scholars to chat with. The visualization helps the user to gain the interact with the recommendation result by the presence of multiple relevance prospects. The high-level characteristics can be summarized as 1) explore the interrelationship between users as well as agents and users. 2) identify relevant items (conference talks) in multiple relevance prospects. 3) provide transparency and increase the trust of the recommended items. The low-level characteristics of this visualization are: to discover a set of recommended items. The user is with no prevalent knowledge of the target and location for an explore action. After the exploration, the user can identify the relevant recommended items. Three types of visualization tasks are proposed: 1) to show the relationship between items associated with different entities. 2) to show clusters of talks linked by connected components. 3) the user can select and confirm the recommended items in a textual list.

The visualization design of *TalkExplorer* can help to improve user understanding of the "black-box" issue behind the recommendation system by providing the controllability and multiple relevance relationships. The user can select the entities from the tag, user, and recommended agents that to be displayed in a clustering style map with related links. The user can further confirm or explore the detail by clicking the visualized clustering or entities through a textual ranking list. The main contribution of this paper can be summarized as 1) the system provides visualization to helps users to efficiently and correctly explore the

conference talks in multiple relevance prospects, which 's hard on a separate page or tables. 2) the visualization helps the users to understand the reason for getting the recommended item by revealing the relationship between data and algorithm. The user can directly inspect the clustered recommended item from different agents or users.

Second, Parra et al. [104] proposed *SetFusion*, a user-controllable transparent hybrid recommendation interface with the CN3 system. The interface adopted a Venn diagram to show the item interrelationship of multiple recommender algorithms. The goal of the system is to provide explanations on a paper recommender system for improving the aspects of privileged transparency, scrutability, and user satisfaction. The target user is the conference attendee who needs to filter the information for the useful paper of the conference. The visualized interface helps the user to gain more controllability and inspectability on the recommended items from multiple recommended methods. The high-level characteristics can be summarized as 1) explore the recommended item interrelationship between a fusion of weights and recommended methods. 2) identify relevant items (conference papers) in a hybrid recommender system. 3) provide controllability on the fusion of methods and the inspectability on the recommended item list. The low-level characteristics of this visualization are: to discover a set of recommended items. The user is with no prevalent knowledge of the target and location for an explore action. After the exploration, the user can identify the relevant recommended items. Based on the goal, three visualization task was proposed, included 1) the list of recommended items: presents recommended papers ranked by relevance from high to low. A color bar was attached to the left side of each paper to indicates the used recommender methods. 2) the weights sliders: use to control the fusion of three recommendation methods. Each slider was associated with the recommendation method with one color code. 3) the Venn diagram: this is the primary visualization task. This is a set-based representation of the recommended item by each method. Three ellipses represent the methods and each paper as the small cycle of the recommended paper. The paper suggested by more than one method was described in the intersection of ellipses.

According to the visualization tasks, the paper visualizes three types of information in the system, including the content of the conference paper, the weights of allocating to each recommender algorithm and the standalone and intersection relevance from three recommender

algorithms. The content of the conference paper is to present the paper title, author, and abstract. The method weights are a linear number from 0 to 1 shows by the sliders with associated color code. The standalone and intersection relevance are displayed on a Venn diagram with three color-coded ellipses and the four intersection areas. The user can explore the paper cycle in a different location for the item recommended by a single method (standalone) or the item recommended by two or more methods (intersection). The visualization design of *SetFusion* provides an intuitive interface for the user to explore the recommended item between one to three recommended methods. The Venn diagram is a widely used visualization to show the intersection relationship between sets of data. The user can gain the transparency of the recommended items from multiple recommended algorithm by the explanation of the diagram, which leads to higher user perception and satisfaction. The user can explore the different combination of methods by changing the slider and confirm the recommended item in the ranking list.

2.3 CONTROLLABILITY IN RECOMMENDER SYSTEMS

Offering users some form of control over the recommendation process can be achieved by two approaches. The first method is through preference elicitation: let the user tell the system what they like, e.g., through forming a user profile [62] or through an adaptive dialog [73]. The second method is through controlling the results: let the user adjust the recommendation profile [55] to fuse recommendations from different sources of relevance [132, 29] or to influence the presented layout [9, 140, 104]. While focused on control, all these approaches contributed to increased transparency of the recommendation process.

Hybrid recommender systems [12] have been gradually becoming more and more popular due to their ability to combine strong features of different recommender approaches. One promising hybridization design is the paralleled hybrid recommender [12], which fuse recommendation results produced by diverse types of existing recommender algorithms as well as multiple kinds of traces left by modern internet users, i.e., browsing trails, bookmarks, ratings, created social links, etc. Typically, paralleled hybrid recommender fuses multiple

relevance sources by assigning static weights to different sources. The optimal weights are trained or learned using ground truth data (i.e., known ratings). The problem with this approach is that users might seek recommendations for different reasons and in different contexts. The individual sources in a hybrid recommender might become more or less valuable depending on each case. As a result, while the "optimal" static fusion could provide the best ranking with high algorithm accuracy, it might be sub-optimal for the users in some specific cases. The problem of optimal source fusion has been originally explored in the domain of information retrieval where it was demonstrated that the user might be in a better position to decide which weight should be assigned to each relevance source in each case [3].

Bringing user control to a hybrid recommender system allows the users to have an immediate effect on the recommendations [62], i.e., the users can further filter or re-sort the recommendation based on their preference or information need. It usually requires an interactive visualization framework that combines recommendations with visualization techniques to support user interaction or intervention into the recommendation process [56]. The idea of the user-controllable interface of different recommendation approaches was originally presented in [113]. Bostandjiev et al. [9] suggested a slider-based interface that the user can adjust the weights of the items and the social connections. Following that, the use of sliders as a way to support user-controlled fusion has been explored in the domain of recommender systems [104] and information retrieval [30] brings additional evidence in favor of using sliders for user-controlled personalization. Verbert et al. [140] encouraged users to choose the most appropriate sources of relevance for each case and provided a cluster-map interface to support user-driven exploration and control of tags, agents, and users. Ekstrand et al. [34] discussed a recommender-switching feature to let the users choose recommender algorithms. Tsai and Brusilovsky [132] offered user-controllable interfaces, a two-dimensional scatter-lot, and multiple relevance sliders, to a social recommender system for conference attendees. Bailey et al. [5] further provides visualization for data analytic tasks using the conference data.

User controllability has also been recognized as a crucial component in supporting the exploratory search, i.e., allowing the users to narrow down the number of items and inspect

the details during the information-seeking process [29]. Ahn et al. [2] presented a summary of search results in the form of entity clouds, which allows the users to explore the results in a controllable interface. Han et al. [54] offered users an option to re-sort people search results based on multiple user-related factors. Di Sciascio et al. [30] proposed a *uRank* interface for understanding, refining and reorganizing documents. [29] integrated controllable social search functionality into an exploratory search system. An effective interactive visualization representation can enable users to control the process of recommendation [56].

2.4 EXPLAINABILITY IN RECOMMENDER SYSTEMS

An alternative approach to increase the process transparency and user satisfaction explored in the literature is providing explanations for recommendations [19]. Explanations that expose the reasoning behind a recommendation could especially increase system transparency [126]. The recommender system is generating recommendations with user-generated data and algorithms. The "reasons" of the receiving the recommendations usually stay in a "black box" [58] that the user has little understanding about the mechanism behind the system. That is, it is a system with low transparency, which has been proved the association with low user satisfaction [126]. To gain the transparency, the study of [126] argued to provide proper explanations on helping the user to understand the related information of the recommending items. That is to give more details that make the users realize the reasons for receiving the recommendations. In a user-centered evaluation, the explanation may significantly contribute to the user experience. The author of [126] summarized seven explanatory goals.

- *Transparency*: the goal is to justify how the recommendation was chosen. In some domains, the transparency of the system is quite important, e.g., the medical decision support system [11]. It is also crucial in increasing usability and user experience, which with a higher user acceptance and preference [119].
- *Scrutability*: the goal is to let a user reflect the incorrectness of the system. The scrutability allows the user to change the "reasoning" of the system, or controls the weighting or

parameter that re-order or re-generate the recommendations. An explanation function may allow the user feedback or control to gain the system scrutability.

- *Trust*: the goal is to increase the confidence in the system. The user may intend to re-use the system due to the trustworthy from the system [18] or an accurate recommendation algorithm [91]. The two mentioned factors can be improved by a high-quality explanation function [78].
- *Persuasiveness*: the goal is to convince the user to try. The explanation can increase the user acceptance of the system suggestions [58], which actively influences the user behavior to utilize the system. A proper explanation may help the user to realize the reasons for receiving the recommendation and then increase the acceptance of the suggestions.
- *Effectiveness*: the goal is to help the users make a good decision. In the decision support system, the system aims to help the user to make the right decisions. For example, in a movie recommender system, the system provides more detail about why the user should select a particular movie. The explanation may help the user make a better decision. In a recent "human in a loop" research, an explanation function can plan a crucial role in helping the user to make the critical decisions [65].
- *Efficiency*: the goal is to help users make a decision faster. Efficiency is one of the factors of system usability. It represents how easy the user can make the concrete decision from the recommendations. It is a "critiquing" process during the utilization [106] across different user preference. A critiquing could be determined by the time used during the search or exploration processes.
- *Satisfaction*: the goal is to make the system enjoyable. It is a critical factor in constituting the user experience. The user may enjoy to use (or re-use) the system if they feel the system usability is high. A high-quality explanation function is positively associated with user satisfaction [41], which is an important factor in a user-centered evaluation framework [71].

Recommender systems explored two principal ways to offer users some form of control over the recommendation process. The first method is through preference elicitation: let the user tell the system what they like, e.g., through forming a user profile [62] or through an adaptive dialog [73]. The second method is through controlling the results: let the user adjust

the recommendation profile [55] to fuse recommendations from different sources of relevance [132, 29] or to influence the presented layout [9, 140, 104]. While focused on control, all these approaches contributed to increased transparency of the recommendation process. An alternative approach to increase the process transparency and user satisfaction explored in the literature is providing explanations for recommendations [19]. Explanations that expose the reasoning behind a recommendation could especially increase system transparency [126]. In an attempt to combine these independent streams of research, I focus on adding explanations to a controllable interactive social recommender interface and study users' subjective feedback and behavior across all design components.

Many scholars have suggested different explanation functions to increase the inspectability of the recommender system [75]. The function provides the transparency that let users realize how the system works [125, 58]. The exposure of the recommendation process through visual interfaces can also increase the inspectability of the system [70]. Many different types of research have been done on this subject. For example, Tsai and Brusilovsky [131] provides recommendation visualization to increase the transparency of the recommender system. Verbert et al. [140] provides a set-based visualization to let the user explore the desired recommendation items. Other researchers further indicated that the value of explaining interfaces could enhance user experiences. The explanation interface was associated with the perception of recommendation quality [125], gaining trust in the system [24] and experiencing the competence of the system [144]. The studies of [45, 98, 97] have mentioned that providing a controllable interface in the social recommender system can increase overall user satisfaction. The authors adopted interactive graphical interfaces to present the social recommendations that enable the controllability of an item or user-level preference in a collaborative recommender system.

Enhancing explainability in recommender systems has drawn more and more attention. Explaining recommendations can achieve different *explanatory goals* by single-style or hybrid explanations [125, 116, 75]. A number of explanation interfaces and approaches have been proposed and studied to assess the improvement of user satisfaction and other aspects [125]. However, most of the evaluation focus on solely user perception or preferences [103, 75]. In most cases, it remains unclear whether different kinds of explanations could improve the

objective parameters of user performance rather than their perception of which option is better while *inspecting* the explanation interface [70]. In recent years, researchers in the field of recommender systems explored a range of advanced interfaces to support exploration, transparency, explainability, and controllability of recommendations [56].

Controllability enabled end-users to participate in the recommendation process by providing various kinds of input [9, 104, 131], e.g., adjust preference or explore recommendations. *Transparency* features allowed interactive recommender systems to deal with the "black-box" problem, i.e., to explain the inner logic of the recommendation process to the end users [120, 141]. A visual interface for the user-controlled hybrid fusion of recommender sources cannot assure that the users will understand the underlying rationale of each contributing recommender; namely, the recommendation algorithm [62]. In the case when a recommendation mechanism is too complicated for non-professional users to explain, some considerable transparency could be achieved by *explainability*, i.e., the system may just need to *justify* why the recommendation was presented [124, 132, 29].

I believe that to increase the transparency of social recommender systems, and interactive user interfaces should be augmented with multiple kinds of explanations for each recommendation source or engine [41, 75]. For example, Papadimitriou et al. [103] proposed a three-dimensional explanation model using human, feature, and item information for explaining social recommendations. A useful explanation model would help users to understand the recommendation reasoning process, which allows the users to make a better decision or persuade them to accept the suggestions from a system [125]. Nonetheless, little is known about how the user will interact with the system when both the fusion process and reasoning process are transparent.

2.5 USER-CENTRED DESIGN AND EVALUATIONS

According to the literature review, there is two mainstream of evaluating social recommender system (RS) with beyond relevance factors. The first stream is the *offline experiment* to test the effect of applying the beyond relevance factors to RS, using an existing data set. The

standard approach is to conduct cross-validation for algorithm efficient or learning-to-rank for ranking performance. The second stream is conducting *human subject study* to see if the user-perceived or satisfied the recommendation system through the interface or modeling algorithm. The two mainstreams will be discussed in this section.

- *Offline Experiment*

The offline analysis is one stream of the evaluation of beyond-relevance factors. Avoiding the trade-off between accuracy and beyond relevance factors is hard. In this review, I found the literature mentions the offline evaluation in two ways: 1) test with proposed metrics or 2) test at a fixed accuracy level. For example, Moody et al. [**?**] divide the Top-N recommendation into different quadrants. The adopted the entropy metric to measure the diversity of a list of recommendation results. Vargas et al. [139] proposed learning to rank evaluation framework with novelty and diversity metrics. The idea is to provide configurations on the rank and relevance of diversity and novelty metrics. The novelty metric is defined as:

$$Novelty = EPC = C \sum_{i_k \in R} disc(k) p(rel|i_k, u)(1 - p(seen|i_k)) \tag{2.1}$$

where $disc(k)$ represents the parameter on ranking position, $p(rel|i_k, u)$ represents the parameter on item relevance. $1 - p(seen|i_k)$ is the item popularity which measures the frequency of viewing by other users.

The article of [110] adopted the SVD with 50 features, included accurate, novelty, and diversity. By an offline music dataset, the goal of the experiment is to generate the most precise recommendation list. The analysis showed a best-recommended model could be combined with different features in each objective, but not all of the objectives. This study adopted the recall, precision, and the EPC model to test the model performance.

Pampin et al. [101] analyzed the performance of item-based and user-based k-NN approaches in quality factors of accuracy and diversity. The authors used a MovieLens dataset with a different ratio of the user and item-based approaches. The experiment result showed a user-based approach generates more different results than the item-based approach. They propose many useful metrics to evaluate neighborhood-based recommender systems. For

example, the metric of 1) Diversity: pairwise comparison of the items in the list by cosine similarity. 2) Popularity: based on the rating for the item and the total number of ratings for all items in the system. 3) Uniqueness: based on the difference between the two recommendation list generated by various algorithms. 4) Precision: based on the intersection of recommended and reinvent items.

Bellogin et al. [6] conducted a study to compare three recommendation approaches: rating-based, content-based, and social techniques. The experiment was used three famous offline datasets for different approaches with a diversity-enhanced metric, e.g., α-nDCG [22] plus the accuracy metric like NDCG. The test result showed the beyond-relevance factors were shown in a different dataset with different methods. This result showed the beyond-accuracy was varied between the dataset. This limitation leads to the second stream of the evaluation approach.

- *Human Subject Study*

It is necessary to collect user feedback to evaluate the factor of beyond. Some literature provided the framework for assessing the user feedback on RS. Pu et al. [108] proposed a study to determine a set of recommendation quality criteria of a user's perception of the usefulness of the system. They classified the rules like 1) Perceived Accuracy: the degree of the recommended items match the user's preference and interests. 2) Novelty: the user receives new, and interest suggested items. 3) Attractiveness: the recommended items were attracted the real desire and attraction. 4) Diversity: the users were not bounded by the same set of recommended items. 5) Context compatibility: consider the context features to provide a context-aware recommendation. The authors conducted a correlation analysis of the proposed 32 criteria within 15 categories. The result showed user-perceived higher satisfaction by the perceived accuracy and novelty.

Knijnenburg et al. [71] proposed a framework for evaluating users' experience of recommender systems. The framework included a factor set of accuracy, satisfaction, choice difficulty, and diversity. The authors surveyed to collect the subject feedback from participants. The study result showed when users perceived the diverse of the recommendation list. It is with a positive relationship with perceived accuracy and eventually leads to higher

user satisfaction. Ekstrand et al. [33] followed the framework to conduct a user study state-of-the-art recommendation algorithms. The experiment result showed the user preferred on the SVD item-item algorithms, but not the user-user algorithms.

Ziegler et al. [153] evaluated the recommendation diversification through a user study. The user was randomly assigned to a user-based and item-based CF recommender. Each of the users was asked to rate the relevance, diversity, and overall satisfaction. The experiment uses a slight diversification on item-based CF increased user satisfaction, but not in user-based CF. Celma et al. [15] conducted a similar study on music recommendation. The author asked the users to rate the familiarity and appreciation of the songs recommended by three algorithms. The experiment showed a complementing pattern between novelty and accuracy metrics.

Hu et al. [60] proposed a visual interface to display the category diversity, compared to the ranking list. They adopted the interface from an online shopping site with an organization interface for a user to browse the related products. The user study by questionnaire showed the user did perceive the product diversity and with higher user satisfaction. A similar visual discovery study was conducted by [122]. They used the Click-through-Rate to examine if the users browse a more diverse set of products.

Willemsen et al. [146] considered a user study on three levels of diversity - low, medium, and high. For each level, the perceived diversity and attractiveness were measured. The experiment result indicated the user did perceive high diversity in the high degree of diversity, but not for the attractiveness factor, which means the user may not appreciate a recommendation list with variety. A similar study was done by [39] to put the "diverse item" in a different position of the recommendation list. The pilot study showed the users were interested in the extra information about the diversity. Castagnos et al. [14] showed a user survey result for the users' appreciation of the transparency of the recommended items. Zhang et al. [152] proposed a music recommendation system and offered a user study to ask the participants to record the familiar, enjoyable, and serendipity. The serendipity-enhanced interface was with high user preference compared to the baseline, although it is less pleasant than the accuracy-oriented interface.

One advantage of human subject experiments is to explain user experience. An efficient

recommender system usually consists of multiple facet components, i.e., user interface, interaction mechanism, algorithm, or even aesthetic. A survey of a direct question may not always reflect the full user feedback. For example, a survey question of *"I am satisfied with the system."* with five or seven scales is a common question to collect the user feedback of system satisfaction. But it is not clear how should the researcher to interpret the feedback to a particular system aspect. The user may feel the satisfaction of the system due to a user-friendly interface, no matter how efficient the recommendation algorithm is. If a researcher considers this as the complete user feedback and uses it as the evidence of an efficient algorithm, the result is biased by the unobserved variables.

It is challenging to "explain" the user experience for the proposed diversity enhanced interface design. Knijnenburg et al. [71] proposed a user-centric evaluation framework for recommender systems in explaining the user experience. The framework is structured with two parts. First, the measured latent concepts should be examined through exploratory factor analysis (EFA), so a researcher can confirm the latent concepts are associated with a certain conceptual component in the framework. Second, the research needs to test the structural relations between the manipulations (system aspects), latent concepts, and behavioral measurement [72]. For the recommendation system evaluation, the author proposed a framework represents six interrelated conceptual components, which can extensively answer the meditating effects beyond the objective and subjective aspects.

1. **Objective System Aspects (OSAs):** As a recommender system is typically multi-faceted with algorithms and interface designs, it is required to isolate the subset of all system aspect in each experiment, so it is possible to claim the effects between control measurements and the other measurements. Hence, in this framework, Knijnenburg et al. [71] defined the objective system aspects as the system aspects that are currently being evaluated, for example, the algorithm, number of recommendations, interface design or other interactions mechanisms. In this dissertation, the OSA represents the diversity-enhanced interface with different manipulations.
2. **Subjective System Aspects (SSA):** The ultimate goal of this framework is to explain the user experience among different objective system aspects. However, even the single system aspect is isolated for testing; the users still need to interact with the mul-

tifaceted system. The user experience should be incrementally increased through the controlled and manipulated aspects. For instance, if an experiment manipulates the recommendation algorithm as the objective system aspect and remains the interface design as controlled aspects, the key is to measure if the user can **perceive** the manipulation so that it can contribute to the overall user experience. Hence, Knijnenburg et al. [71] proposed the subjective system aspects as the mediating variables of user experience, user interactions, and the objective system aspects. The variables are the moderators that help to establish connections through user perception on certain system aspects.

3. **User Experience (EXP):** The user experience is a self-relevant subjective metric that reflects the emotions and attitudes of a user while using the system. Knijnenburg et al. [71] classified the user experience into three types. First, the system-related user experience is measured user perception of the system's effectiveness. Second, the process-related user experience that determines if the user can choose or browse the recommendation efficiently. Third, the outcome-related user experience measures if the recommendations can help to decrease choice difficulties. The three types of user experience, which comes with multiple constructed survey questions, can help researchers to understand and explain the overall user experience on different system aspects.
4. **User Interaction (INT):** One of the crucial aspects of using the recommender system is user interactions. There are two kinds of approaches to measuring user interactions: subjective or objective. The subjective method is based on the user feedback regarding the interactions while using the system. For example, the users may be questioned on the intention of (re-)using the system or rating the recommendations. The objective approach used the logged data, i.e., the number of recommendations clicked and inspected by the user or the time they spent on using the system. The user interactions help to explain and understand the effects of SSA and EXP above.
5. **Personal and Situational Characteristics (PCs and SCs):** The final two components are related to the user instead of the system aspect. These two are used to test the influence of the user's characteristics and situational awareness while using the system. The factors are beyond system aspects but have a significant impact on EXPs. It is essential to consider the difference in the personal characteristics, e.g., the acceptance of

the privacy and the expertise of using the system. Personal aspects usually play a crucial role in subjective feedback, e.g., a user who with better knowledge of the system may have a better chance to realize the technical details of the system. It may contribute to a better trust of the advanced interface design. On the other hand, the situational characteristics are worth considering while explaining the user experience. For example, does the user really interact (or realize) the function in the system? In many cases, it is not surprising that the users "ignore" the new design functions. A post-experiment survey can help to measure if the user realizes or adopts the specific system components.

3.0 RESEARCH PLATFORM AND EXPERIMENT SETTINGS

This chapter presents the research platform and the shared experiment settings in my dissertation. I first introduce the social recommender system *Conference Navigator* and then continue with the detail of the recommendation models, the measurements of recommendation quality and diversity. The research platform and experiment settings were adopted across all studies in this dissertation.

3.1 RESEARCH PLATFORM: CONFERENCE NAVIGATOR

Conference Navigator [1] (shown in Figure 1) is a online conference support system, which has been developed to third version [105] and adopted by many different research and studies [140, 10, 104, 127, 130, 131, 128]. The system is with tools to help the conference attendees in browsing the conference program, publication, author, and attendees. *Conference Navigator* system has been used to support more than 45 conferences at the time of writing this paper and has data on approximately 7,045 articles presented at these conferences; 13,055 authors; 7,407 attendees; 32,461 bookmarks; and 1,565 social connections.

The users can follow or connect with scholars based on their interests (shown in Figure 2). The following function is a one-way relation without confirmation from the target user. The connect function requires a confirmation to establish mutual relations. The user can use the functions for different information-seeking scenarios, e.g., to pay attention to a scholar with similar interests or to find a prospect collaboration in the conference venue. One of the major functions is *social recommendation* for filtering relevant authors or attendees by

[1] http://halley.exp.sis.pitt.edu/cn3/portalindex.php

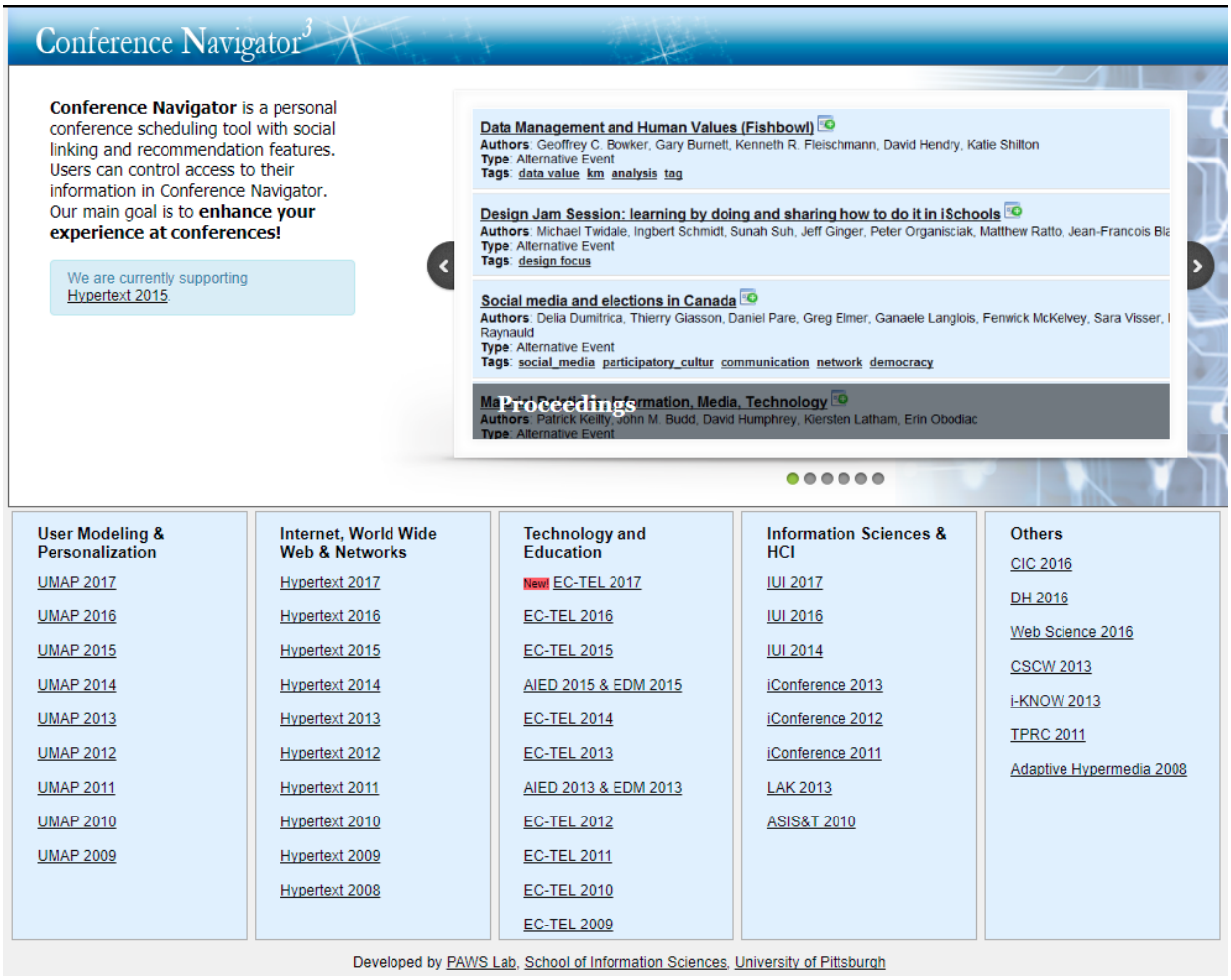
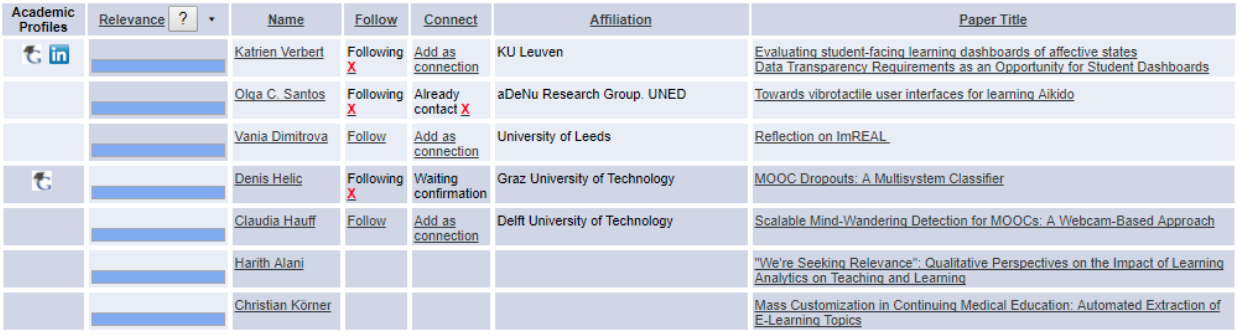


Figure 1: Screenshot of the Conference Navigator System's front page.



Academic Profiles	Relevance ?	Name	Follow	Connect	Affiliation	Paper Title
		Katrien Verbert	Following X	Add as connection	KU Leuven	Evaluating student-facing learning dashboards of affective states Data Transparency Requirements as an Opportunity for Student Dashboards
		Olga C. Santos	Following X	Already contact X	aDeNu Research Group. UNED	Towards vibrotactile user interfaces for learning Aikido
		Vania Dimitrova	Follow	Add as connection	University of Leeds	Reflection on ImREAL
		Denis Helic	Following X	Waiting confirmation	Graz University of Technology	MOOC Dropouts: A Multisystem Classifier
		Claudia Hauff	Follow	Add as connection	Delft University of Technology	Scalable Mind-Wandering Detection for MOOCs: A Webcam-Based Approach
		Harith Alani				"We're Seeking Relevance": Qualitative Perspectives on the Impact of Learning Analytics on Teaching and Learning
		Christian Körner				Mass Customization in Continuing Medical Education: Automated Extraction of E-Learning Topics

Figure 2: Screenshot of Conference Navigator System's social recommendation page.

fusing multi-relevance data. The function is aimed to help the users to facilitate social interactions better. However, it also creates a challenge in providing the recommendations in a hybrid model that fused multiple relevances. This dissertation presents my exploration of how to bring controllability and explainability to a hybrid social recommender system in the *Conference Navigator* system.

3.2 RECOMMENDATION MODELS

The hybrid social recommender system will rank the recommended attendees by their relevance to the target user. It is a content-based recommender system that personalizes the information to a user based on users' interests or relevance. The system uses five separate recommender engines (models) that rank other attendees along five dimensions. I select the similarity measure through the nature of each recommendation model. First, the *Publication Similarity* is calculated by the text-similarity of their academic publication text. I use *cosine similarity* to measure the similarity between two termvectors. It is a common measure for comparing the similarity between two documents in the area of text mining. Second, *Topic Similarity* is calculated by their research interests. I use *Jaccard similarity* to measure the similarity between sets of topical words, which are generated by the topic modeling approach. Third, the *Co-authorship Similarity* is the overlap and distance of the co-authorship network. I measure the similarity through network distance and overlaps. Fourth, the *CN3 Interest Similarity* is the similarity of their bookmarks in the Conference Navigator system. I use *Jaccard similarity* to measure the similarity of two sets of bookmarked items. Fifth, the *Geographic Distance* is representing the distance between the user's affiliation. I use an ad-hoc approach to measure the geo-distance between two locations.

The recommendation models are discussed as below:

1. **Publication Similarity** is determined by the degree of publication similarity between two attendees using cosine similarity [89, 138]. The function is defined as:

$$Sim_{Publication}(x, y) = (t_x \cdot t_y)/\|t_x\|\|t_y\| \tag{3.1}$$

where t is word vectors for user x and y. I used TF-IDF (Term FrequencyInverse Document Frequency) to create the vector with a word frequency upper bound of 0.5 and lower bound of 0.01 to eliminate both common and rarely used words. The TF-IDF method is widely used in information retrieval systems and a content-based recommender system. The formula is shown below:

$$tf(t,d) = ft,d \tag{3.2}$$

$$idf(t,D) = log\frac{N}{|d \in D : t \in d|} \tag{3.3}$$

where the t represents the word, d represents a certain document and D represents a set of documents. "tf" stands for the frequency of a word in a document. "idf" represents the inverse of the document frequency among the whole corpus of documents. The purpose of tf-idf is to highlight the importance of a certain word in a document. For example, if one word appears in all documents, it may refer to a preposition which with no actual meaning. So I choose a ratio from 0.01 to 0.5 to eliminate both common and rarely used words.

2. **Topic Similarity** is a metric that measures the Distance between topic distributions [31]. The approach assumes that a mixture of topics is used to generate a string (document), where each topic is a distribution of topical words. In my dissertation, the *topics* were generated by topic modeling, Latent Dirichlet Allocation (LDA), by classifying their publication text [31]. A higher topic similarity means a shorter distance between the two scholars' research interests, i.e., the two scholars shared more common research topics.
3. **Co-authorship Similarity** approximates the social Similarity between the target and recommended users by combining co-authorship network distance and common neighbor similarity from published data. In pre-study, I adopted the depth-first search (DFS) method to calculate the shortest path p [121] and common neighborhood (CN) [95] for the number n of coauthors overlapping in two degrees for user x and y.

$$Sim_{Co-authorship}(x,y) = p + n \tag{3.4}$$

Depth-first search (DFS) is an algorithm for traversing or searching tree or graph data structures. I plan to formalize the co-authorship as a graph. The shortest distance will determine the DFS from the original user to the target user. It is a method to measure how close the two scholars link to each other. The formula is shown as: Let $G = (V, E)$ be a graph with n vertices of V. For $\alpha = (v_1, ..., v_m)$ be a list of distinct elements of V, for $v \in V(v_1, ..., v_m)$, let $v_\alpha(v)$ be the greatest i such that v_i is a neighbor of v, if such i exists or be 0 otherwise.

The common neighborhood (CN) [95] indicates the intersection set of neighbors of a given author. Here I define the set of neighbors as all co-authors observed at t. The formula is shown as: Let $G = (V, E)$ be a graph with n vertices of V. For $(v_1, ..., v_m)$ be a list of distinct elements of V. The common neighborhood graph (congraph) of G is a graph with vertex set $(v_1, ..., v_m)$ in which two vertices are adjacent if they have at least one common neighbor in the graph G. The formula will return the total number of common neighborhood or 0 otherwise. I consider only the one-degree relationship, which is also possible to extend to more degrees based on the system's needs.

In study 5-6, I further extend the method to Personalized Hitting Time [85]. The method adopted the theory of random walk, which provides a more sophisticated performance in ranking the recommendations. Assuming given a weighted digraph G, let $(x_t)_t >= 0$ be a standard random walk on G. Define the random variable $\rho_j = intt : X_t = j$. The hitting time between two nodes i and j is

$$Sim_{HittingTime}(i, j) = \mathcal{E}(\rho_j | X_0 = i) \tag{3.5}$$

4. **The CN3 Interest Similarity** is determined by the number of co-bookmarked papers and co-connected authors within the experimental social system [10]. The function is defined as

$$Sim_{CN3}(x, y) = (b_x) \cap (b_y) + (c_x) \cap (c_y) \tag{3.6}$$

where b_x, b_y represent the paper bookmarking of user x and y; c_x, c_y represents the friend connection of user x and y. The intersection is calculated by Jaccard Coefficient.

The Jaccard Coefficient (JC) [21] measures similarity between finite neighbor sets. Here I defined neighbors sets as co-bookmark or co-connection sets at t. For any two given authors, it is the intersection of their co-authors sets divided by the union of their co-authors sets. It is computed as $Sim_{JC} = \|\Gamma(x) \cap \Gamma(y)\| / \|\Gamma(x) \cup \Gamma(y)\|$, where x or y is the given author and $\Gamma(\cdot)$ represents the co-bookmark or co-connection they have.

5. **The Geographic Distance** is a measure of geographic Distance between attendees. I retrieve longitude and latitude data based on attendees' affiliation information. I used the Haversine formula to compute the geographic Distance between any pair of attendees [138].

$$Sim_{Distance}(x, y) = Haversine(Geo_x, Geo_y) \tag{3.7}$$

where *Geo* are pairs of latitude and longitude coordinates for user x and y, the *Geo* information is determined by the users' affiliation data. For instance, for a scholar who comes from the *University of Pittsburgh*, the latitude and longitude coordinate as $(40.440625, -79.995886)$. I use Google Map API to convert the affiliation information (city, country) to the latitude and longitude format.

The Haversine formula can be used to calculate any two points on a sphere,: gives the Haversine of the central angle between them.

$$hav(\frac{d}{r}) = hav(\rho_2 - \rho_1) + cos(\rho_1)cos(\rho_2)hav(\lambda_2 - \lambda_1)) \tag{3.8}$$

where hav is Haversine function stands for $hav(\theta) = sin^2(\frac{\theta}{2}) = \frac{1-cos(\theta)}{2}$. d is the distance between the two points (along a great circle of the sphere), r is the radius of the sphere. ρ_1, ρ_2 are latitude of point 1 and latitude of point 2, in radians. λ_1, λ_2 are longitude of point 1 and longitude of point 2, in radians.

3.3 EXPERIMENTAL MEASUREMENTS

3.3.1 Recommendation Quality

In this dissertation, I adopted the following measurement for recommendation quality.

- **TopN@K:** The metric measures the accuracy of a list of k recommendations [88]. For a user u, the TopN@K of a list of recommendation is

$$TopN@k = \frac{|rel|}{k} \tag{3.9}$$

- **Mean Reciprocal Rank (MRR):** The metric measures a list of possible responses to a queries [25]. The queries ordered by probability of correctness, which is measured by an inverse value of the rank of the first correct answer, e.g., score $\frac{1}{2}$ for the query of correct answer shown in second place .

$$MRR = \frac{1}{k}\sum_{i=1}^{|k|}\frac{1}{rank_i} \tag{3.10}$$

  where $rank_i$ is the position of the first correct answer (relevant item) for i-th query.

- **Normalized Discounted Cumulative Gain (nDCG):** The metric measure the quality of a list of recommendations, which considered the ranking of relevance recommendations [88]. A higher nDCG value means the recommendation better fitting the user preference in top-ranking positions. Perfect ranked recommendations would lead to nDCG metric equal to 1.

$$nDCG = \frac{DiscountedCumulativeGain(DCG)}{idealDiscountedCumulativeGain(IDCG)} \tag{3.11}$$

  where $IDCG$ is a perfect case of the query, which is the highest possible nDCG value. DCG is defined as

$$DCG = \sum_{i=1}^{|k|}\frac{2^{rel_i}-1}{log_2(i+1)} \tag{3.12}$$

  where k is the number of recommendation, rel_i is the relevance score from 0 to 1.

3.3.2 Recommendation Diversity

The system allows the user to rank and visualize items using on different aspects of relevance through our proposed interface. The recommendation diversity can be measured by two diversification models.

1. **Feature Diversification:** the user can select any two pairs of proposed features and spot the recommended items from the intersection of their relevance. All of the proposed features were calculated on a different scale. For example, the distance feature is the physical distance in miles, while the academic feature is calculated as a percentage. To enable the comparison of diverse features, A standard Z-score was adopted to normalize all the features to the same scale, from 0 to 1. The function was defined as:

$$ZScore = \frac{x_i - u_j}{\sigma_j} \tag{3.13}$$

   where x_i is the ith recommended item and j represents the corresponding feature with its average u and variance σ. Then, a standard Z-table was used to convert the $ZScore$ to the corresponding percentile p_{ij}. Hence, all the features can be presented on the same scale, both in a ranked list and scatter plot diagram.

2. **Category Diversification:** it is a model of diversifying the different categories [63]. For example, in the scatter plot of pre-study, I color-code the items from different categories, such as title, position, and country. In the ranked list, I listed the category as one column for a user to access.

3. **Shannon Entropy:** I can then measure the user selection/exploration diversity, based on the two diversification models. I observe the user's interaction with items from different "quadrants" (feature intersections) [127], such as high academic and high social features, or high academic and low social features. The extent of diversity is measured

by *Shannon Entropy*:

$$Entropy : d_u = -\sum_{i=1}^{4} p_i log_4 p_i \tag{3.14}$$

where p_i is the probability for a particular quadrant (feature or category) and the proportion of all of the user's selections [94], based on the definition, I can measure the diversity in the different aspects of the relevance dimension. I can compare the combinations of all the proposed features, for example, in a recommendation system fused with $four$ features. I can measure the entropy difference among the $4 * (4 - 1) = 12$ pair of dimensions.

4. **alpha NDCG:** it is relevant to NDCG but as used to measure for diversified search, where it is appreciated by the number of covered intents [22].

$$alphanDCG = \frac{DiscountedCumulativeGain(DCG)}{IdealDiscountedCumulativeGain(IDCG)} \tag{3.15}$$

where $IDCG$ is a perfect case of the query, which is the highest possible alpha nDCG value. DCG is defined as

$$DCG = \sum_{j=1}^{|k|} J(d_k, i)(1 - \alpha)^{r_{i,k-1}} / (\log_2 1 + j) \tag{3.16}$$

where J represents the intent probabilities of the given iter, $\alpha = 0.5$ is the factor to control the level of diversification, log is the discount function of ranking.

4.0 USER CONTROLLABLE INTERFACES FOR A HYBRID SOCIAL RECOMMENDER SYSTEM

This chapter presents the experiment results of pre-study that helps me to choose an effective user controllable interface for later studies. In this pre-study, I extended the user interface designs from the early two pilot studies and evaluated the design with a larger-scale and real-world conference setting [129, 130]. I present two of my attempts that bring a two-dimension scatter plot (*Scatter Viz*) and ranking-based multi-relevance sliders (*Relevance Tuner*) to a social recommender of academic conferences. The finding indicated a different usage pattern in the two user interface and the *Relevance Tuner* was shown useful in enhancing recommendation diversity as well as receiving positive user feedback. Hence, I will adopt the design of *Relevance Tuner* as the core user interfaces in my study 1 to 6.

4.1 INTRODUCTION

The pre-study presented in this chapter reports my exploration of two visual recommender interfaces. First, I proposed a recommender interface that explores the value of a two-dimensional scatter-plot visualization to present recommendations with several dimensions of relevance. In the context, the scatter plot interface was used to help users combine different aspects of relevance for recommended items while providing inspectability to the users. Second, I proposed a recommender interface that enhances the fusion control function within a ranked list with meaningful visual encoding for multiple dimensions of relevance. The users can adjust the relevance weightings to customize the recommendation results, which provides the user with a higher level of control over their results.

The two interfaces were designed to explore the value of user-controllable and diversity-aware interfaces in a social recommender system. Each of the interfaces has been evaluated in a controlled field study in the target context. The results show that the new visual interfaces reduce exploration efforts for a set of realistic tasks, and also make the users more aware of the diversity of recommended items. Also, the users' subjective evaluation shows a significant improvement in many user-centric metrics. I further discussed the effects of the proposed interfaces on the users' experience with a diversity-enhanced social recommender system.

This chapter offers several contributions: 1) I propose two interfaces that support the continuously controlled fusion of several relevance aspects with inspectability and controllability; 2) I provide evidence that the diversity-aware interface not only helps the user to perceive diversity but also helps the user to improve usability in the real world beyond simple relevance tasks; 3) the experiment results helps me to choose an effective user controllable interface for this dissertation.

4.2 CANDIDATE INTERFACE #1: SCATTER VIZ

In a hybrid recommendation context with multiple types of relevance, the traditional ranked list makes it hard for the user to recognize how different relevance aspects are correlated. A typical example of this situation is recommending other attendees to meet at a research conference. Here a range of similarity functions (social, past publications, current interest, location) could indicate a person worth to meet. To help conference attendees in their conference networking, I proposed a dual social recommender interface, **Scatter Viz**, which includes a ranked list and visual scatter plot components. The ranked list was selected as a traditional way of presenting recommended results in a single dimension, listed from high to low relevance. The scatter plot was chosen as an intuitive way to present multidimensional data [66] and inspect patterns in large data-sets [27], as it has been shown that users can accurately judge data similarities between different shapes of scatter plots [102]. I hoped that the ability to view recommended items in two dimensions could reveal the overall diversity

of results and help to correlate multiple types of relevance among social recommendations.

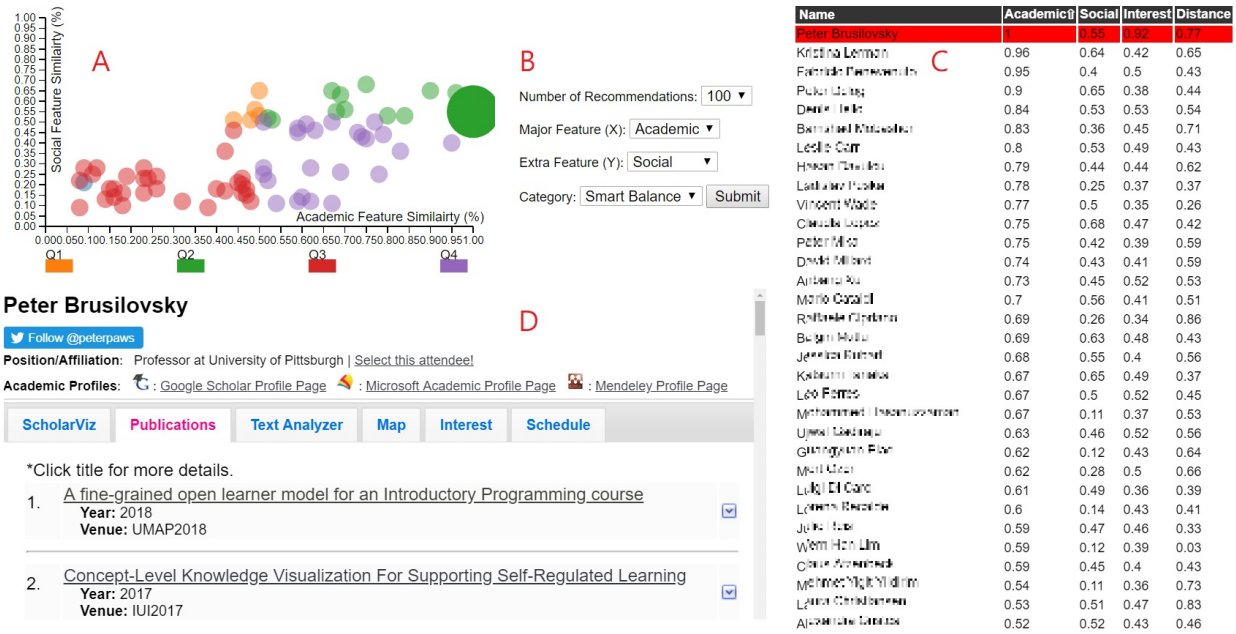


Figure 3: The design of Scatter Viz: (A) Scatter Plot; (B) Control Panel; (C) Ranked List; (D) User Profile Page. The interface supports exploration of recommended items in Section A or C and detail inspection in section D. The scholar names have been pixelated for privacy protection.

Figure 3 illustrates the design of the dual interface in four sections.

1. **Section A** is the scatter plot. The interface presents each item (a conference attendee) as a circle on the canvas in two selected dimensions. The user can move the mouse over the circle to highlight the selection.

2. **Section B** shows the control panel, with which the user can interact. The user can select the number of recommendations to display and choose the *major feature* and the *extra feature* to visualize the recommendations on the scatter plot. The major feature is used to rank the results along the X-axis of the Scatter Plot (Section A) and in the ranked list (section C). The extra feature is used to diverse the recommendations in respect to the selected aspect along the Y-axis (thus spreading the results that have similar values

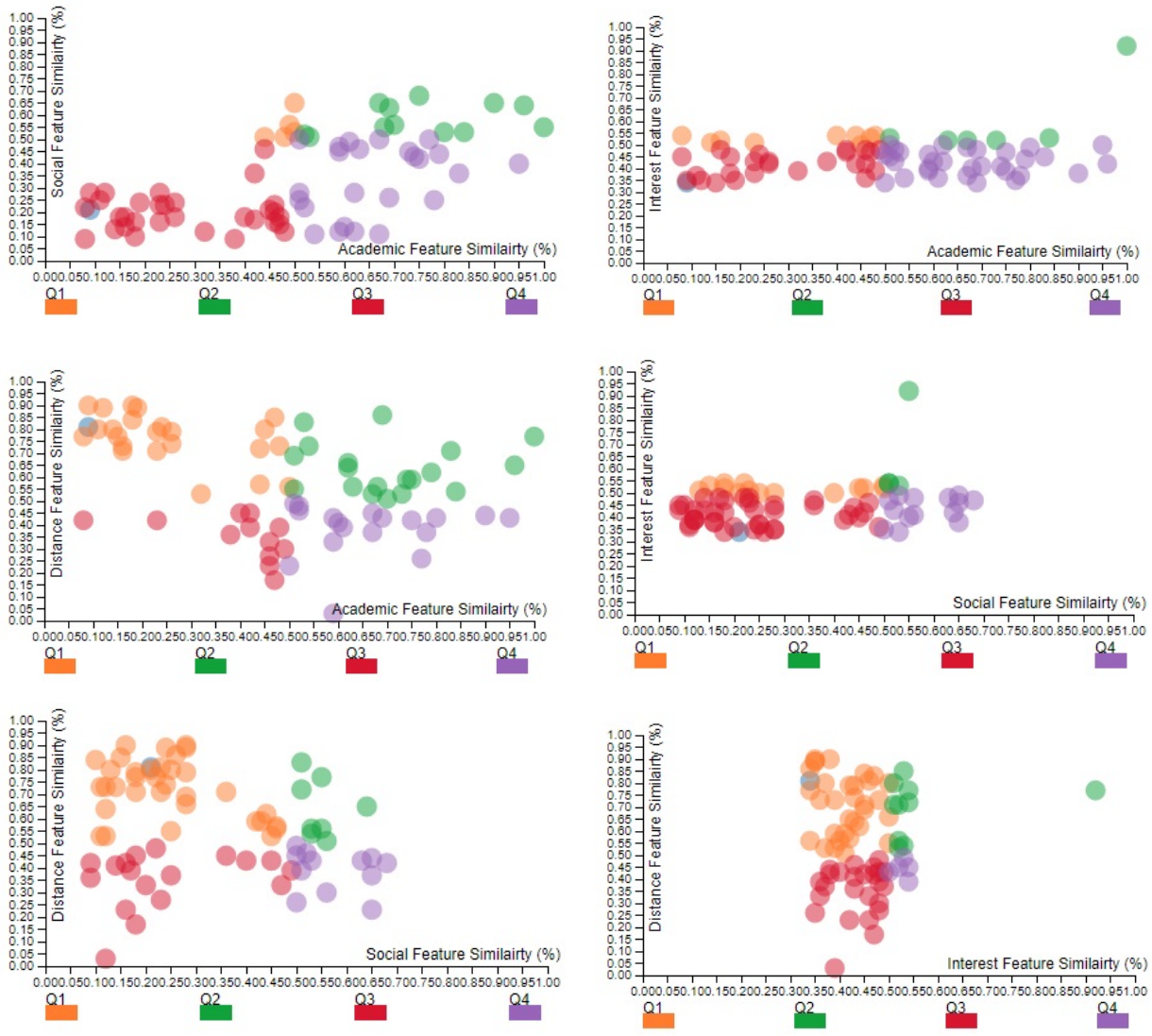


Figure 4: Scatter plot layouts: the layout would adjust, based on the selected *Major* and *Extra* features (Section B in Fig. 3). Here is an example that presents the same recommended items in six dimension combinations: Academic/Social, Academic/Interest, Academic/Distance, Social/Interest, Social/Distance and Interest/Distance feature coordinates. The nodes are colored using four equal quadrants that set Category as *Smart Balance*.

of the major feature but different values of the extra feature). To further investigate the diversity of the displayed recommendations, the user can also use another data aspect

as a *category* to color-code the results. The default category was *Smart Balance*, which highlights four quadrants of the displayed data with a 0.5 ratio. Figure 4 presents six sample scatter plot layout combining "Academic", "Social", "Interest" and "Distance" relevance features that are color-coded using the Smart Balance Category.

3. **Section C** is the standard ranked list. More precisely, it is a combination of four ranked lists produced by four recommender engines, as explained below. To make the four dimensions more comprehensive, the model normalized relevance scores from 0-1 of each user to the target user, generated by each recommender engine. All the relevance scores are shown on the right side of the ranked list. The user can hover over each row to highlight the location in the scatter plot or click for a more detailed user profile.

4. **Section D** presents more detailed information about the person who has been selected in either the visualization or the ranked list. Among other aspects, four of the six tabs visually explain how each recommender engine calculates the relevance of the selected user to the target user. Due to the page limitation, the details of each explanation tabs are omitted. The design detail of the explanation functions can be found in [131].

The visual encoding affects the way users process the information. Pre-attentive processing let users absorb and precept the enormous amount of information in a short period [30]. The proposed interface helps to present the recommendation results in two kinds of visual encoding. First, the interface displays the recommendation relevance in two dimensions. The visual encoding helps the user to *spot* the item in different dimensions. It helps the user to make a decision beyond single or combined relevance, which is more realistic in many real-world scenarios. For example, a user may be interested in a scholar who whose research area is highly relevant to their research and who is also affiliated with nearby cities. The scatter interface helps to filter a group of recommended items with the two desired relevance features. Second, the node is color-coded in different categorical features; for example, in Smart Balance mode, the node is color-coded by the four quadrants between two dimensions of features. The user can perceive the tendencies of the recommendation item, based on their coloring, and the user can also update the layout with different Category features,

including the meta-data of the recommended scholar's title, position, and home country. In addition, both the node and table row are highlighted synchronously while the user moves over the recommended items (see example in Figure 3). As a result, the scatter plot interface can be used for recommended item selection or just as a diversity-oriented recommendation explanation.

4.3 EVALUATION OF SCATTER VIZ

4.3.1 Data and Participants

The recommendations produced by all four engines are mostly based on data collected by the Conference Navigator 3 (CN3) system [10]. The data was using the conference proceeding of the 2017 Intelligent User Interfaces Conference (IUI 2017). A total of 25 participants (13 female) were recruited for the user study. All of the participants were attendees of the IUI 2017 conference. Since the primary goal of the system was to help junior scholars connect with other people in the field, I specifically selected junior scholars, such as graduate students or research assistants. The participants came from 15 different countries; their ages ranged from 20 to 50 (M=37, SE=7.07). All of them could be considered as knowledgeable in the area of the intelligent interface for at least one academic publication from IUI 2017. To control for any prior experience with the recommender system, I included a question about in the background questionnaire. The average answer score was (M=3.28, SE=1.13) on a five-point scale.

4.3.2 Experiment Design and Procedure

To assess the value of the proposed interface, I compared the dual Scatter Viz interface with the scatter plot and the ranked list (I will call this condition as SCATTER) with a baseline interface using only a ranked list (RANK) with Section A (in Figure 3) removed. The study used a within-subjects design. All participants were asked to use each interface for three following tasks and to fill out a post-stage questionnaire at the end of their work

with each interface. At the end of the study, participants were asked to compare interfaces regarding their explicit preference. The order of using interfaces was randomized to control for the effect of ordering. In other words, half of the participants started the study with the SCATTER interface. To minimize the learning effect (becoming familiar with data), I used data from two years of the same conference: the SCATTER interface used papers and attendees from IUI 2017, while the RANK interface used the corresponding data from IUI 2016.

Participants were given the same three tasks for each interface. The tasks were explicitly designed as diverse but realistic tasks that could be naturally pursued by attendees at research conferences.

- **Task 1:** Your Ph.D. adviser has asked you to find four Committee Member candidates for your dissertation defense. You need to find candidates with expertise close to your research field while trying to lower their travel cost to your defense.
- **Task 2:** Your adviser has asked you to meet four attending scholars, preferably from different regions across the world, who have a close connection to your research group.
- **Task 3:** You want to find four junior scholars (not yet faculty members) with reasonably similar interests among the conference attendees to establish networking.

The participants were asked to pick suitable candidates among conference attendees, based on their best judgment in each task. When designing the tasks, I attempted to make them realistic, yet focused on multiple aspects of relevance, as many real tasks are. I consider that task 1 is relevance-oriented and that tasks 2 & 3 are diversity-oriented. For a relevance-oriented task, I expect to see if the proposed interface helps the user to coordinate different relevance aspects of the desired target efficiently. In contrast, for the diversity-oriented task, I expect the system to help to recognize the diversity of recommended items, as compared to the baseline interface.

4.3.3 Action Analysis

Table 1 shows the system usage for two interfaces. The data indicate that participants extensively used both the control panel and explanation tabs to complete the tasks. The

Table 1: User action summary of Scatter Viz: the table shows the user interaction statistics while performing each of the three tasks using two interfaces. (Statistical significance level: (*) $p < 0.05$.)

		RANK		SCATTER		
Task	Action	M (SE)	User Count	M (SE)	User Count	
T1	Control Panel	3.88 (2.40)	24	4.12 (2.02)	25	
	Explanation Tab	34.28 (29.50)	25	7.96 (7.48)	19	
	Click - Rank	26.28 (29.50)	25	4.92 (6.75)	15	*
	Click - Scatter	-	-	3.04 (5.45)	13	*
	Time Spending	345.44 (209.86)	25	389.12 (235.29)	25	
T2	Control Panel	2.88 (1.64)	24	2.88 (1.12)	25	
	Explanation Tab	19.96 (17.47)	25	9.16 (6.28)	25	
	Click - Rank	16.96 (17.47)	25	2.68 (4.69)	15	*
	Click - Scatter	-	-	3.48 (5.41)	13	*
	Time Spending	216.6 (144.95)	25	190.84 (115.33)	25	
T3	Control Panel	2.56 (1.04)	24	2.84 (1.10)	25	
	Explanation Tab	20.08 (20.29)	25	6.4 (7.22)	19	
	Click - Rank	19.08 (20.29)	25	3.48(7.80)	9	*
	Click - Scatter	-	-	2.92 (2.95)	15	*
	Time Spending	345.95 (156.39)	25	369.2 (169.77)	25	

participants usually required more actions on the first task to familiarize themselves with the system. There is no significant difference on the action of change control panel and the click on the explanation tab between the interfaces in three tasks; although, in the SCATTER interface, the users tended to click the explanation functions less. The click frequency presented a significant difference between the two interfaces. This finding is not surprising because the RANK interface lacks the visualization information that pushes the participant to click more on the user profile page to inspect the necessary information. It is

interesting to see that not every user clicked on the scatter plot. This data hints that some participants treated the scatter plot visualization as an explanation function rather than an interactive exploration interface. At the same time, I found no significant difference in the time spent on the tasks. The data hints that each action taken in the SCATTER interface delivered more interesting information with which to engage.

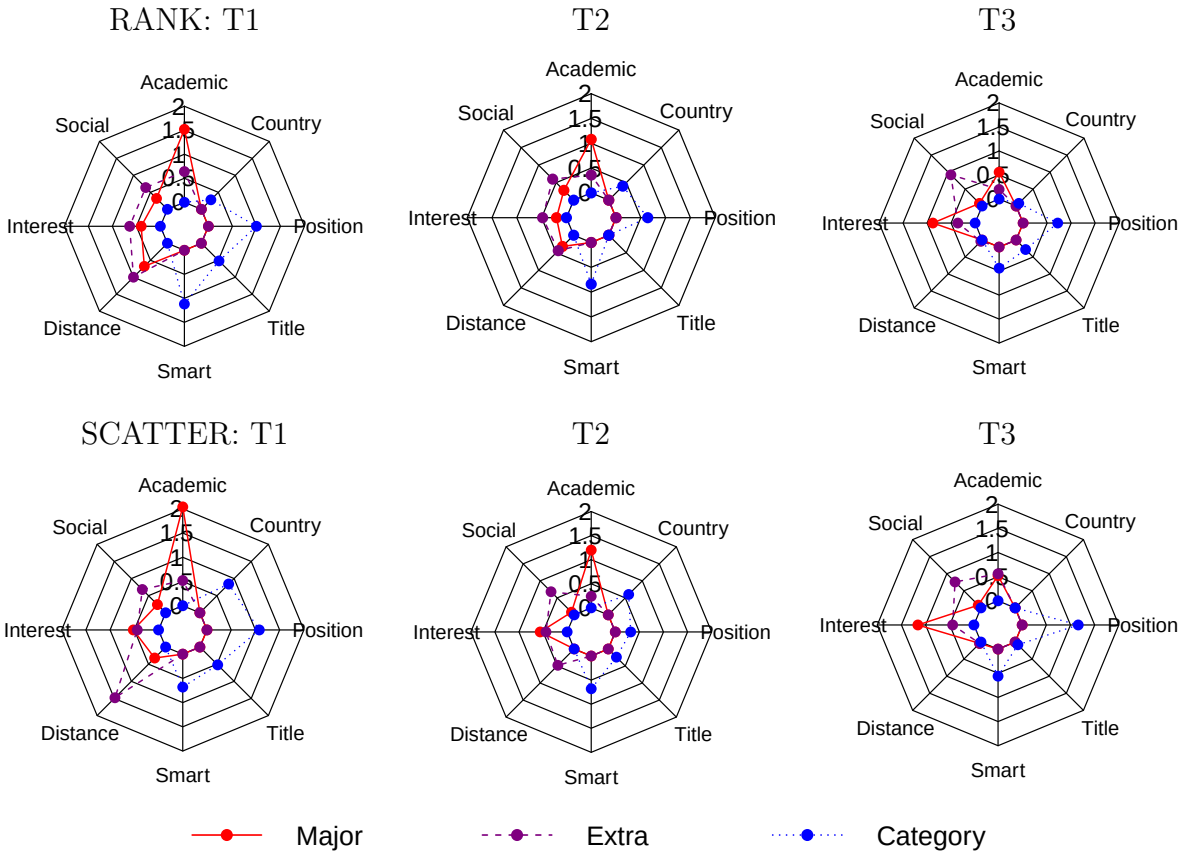


Figure 5: The recommendation features usage of Scatter Viz. The red line is the usage of major features, the purple line is the usage of extra features, and the blue line is the usage of category features.

It would be valuable to see how the participants adopt the control and explanation functions. Figure 5 presents the usage of four similarity features and three category features. The red line shows how frequently the feature was selected as the primary ordering factor. The factor determines 1) the ranking in the RANK interface, and 2) the x-axis layout in the SCATTER interface. The purple line shows the frequency of its use as the second feature to

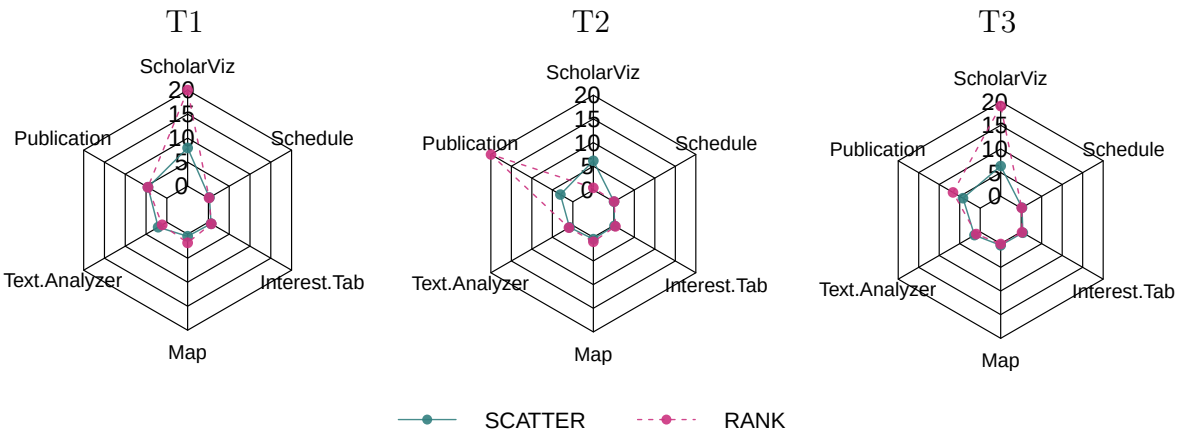


Figure 6: The explanation function usage of Scatter Viz. The light plum color is the usage of the RANK interface, while the light blue color is the usage of the SCATTER interface.

spread the results along the y-axis in the SCATTER interface. The blue line is the frequency of feature used as the Category feature. The chosen category feature creates one additional column in the RANK interface and updates the color-coding in the SCATTER interface. I can observe the usage pattern between the three proposed tasks. In task 1, which focuses on finding the 2-category optimum ranking, the participants most frequently used Academic as the primary ranking feature, diversifying it by Distance and selecting the Position feature to color-code the results. In diversity-oriented task 2, a number of diversification features were tried with about the same frequency. In task 3 (find four junior scholars with similar interests), the Interest feature was most frequently selected as the primary feature, while Social was the primary diversification approach and Position was the primary color-coding approach. The pattern was consistent between two interfaces. Overall, this result shows that the users were quite efficient in selecting the most useful features for each task. Figure 6 shows the click frequency on six different explanation tabs. Overall, the users were most interested with explanations presented in ScholarViz and Publication tabs (which explained the Social and Academic similarities, respectively). The figure also shows that explanations were requested more frequently for candidates accessed using the RANK interface.

Table 2: Post-stage questionnaire [108, 13].

Q1	The interface helps me to explore various interesting people in the conference.
Q2	It is helpful to see people attributes like Title, Country, and Position when exploring interesting people in the list.
Q3	The interface helps me to perceive the diversity of explored attendees
Q4	The interface helps me to improve my trust in the people recommendation result.
Q5	The interface helps me to understand why specific attendees were recommended.
Q6	I like the people recommendation result from the system.
Q7	I became familiar with the system very quickly.
Q8	Overall, I am satisfied with the system.
Q9	I will frequently use the system at future conferences.
Q10	It was useful to see the explanation of scores produced by different recommendation components.
Q11	It is fun to use the system.
Q12	The system has no real benefit for me.

4.3.4 User Feedback Analysis

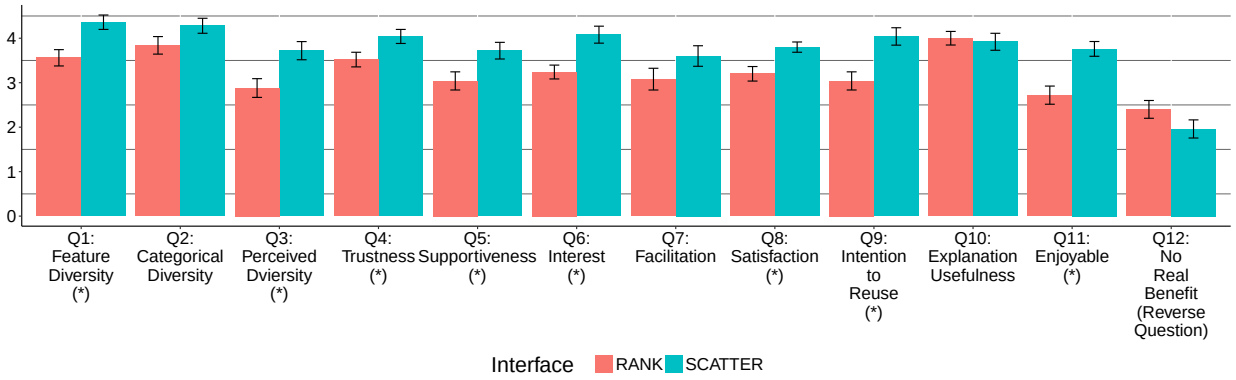


Figure 7: User feedback of Scatter Viz: the result shows that the SCATTER interface received a significantly higher rating for six aspects. (A cut-off value was set at 3.5 on the 5 point scale. Statistical significance level: (*) $p < 0.05$.)

To compare subjective feedback, I analyzed the responses of the post-stage questions using paired sample t-tests. Figure 7 shows the result of this analysis. I compared the twelve aspects of subjective feedback from the participants; among them, the SCATTER interface

received a significantly higher rating for six aspects: Trust (Q4), Supportiveness (Q5), Interest (Q6), Satisfaction (Q8), Intention to Reuse (Q9), and Enjoyable (Q11). In two questions, facilitation (Q7) and the Reversed Benefit Question (Q12), the SCATTER interface scored higher, but not significantly so. It is interesting to see that the RANK interface scored a bit higher (though not significantly so) on explanation usefulness, which hints that the lack of visualization made explanations more important in the RANK interface. In the final preference test, the SCATTER interface received much stronger support than the RANK interface in the user preference feedback (Figure 8). Most importantly, a majority of users (84%) considered the SCATTER interface to be a better system for recommending attendees and better help in diversity-oriented tasks, as well as a better system for recommending.

4.3.5 Recommendation Diversity Analysis

Table 4 shows the diversity analysis for each task and interface. The result shows the users' responses to the tasks with a different pattern of exploration, which caused a variance of diversity and coverage measurements. All three tasks are shown a least one significance between two interfaces but in the different aspects of features. For the SCATTER interface, task 1 (relevance-oriented) shows significance statically on less difference between academic/social & social/interest features, but more coverage on the title category. Tasks 2 & 3 (diversity-oriented) show higher selection diversity in the interest/distance and social/distance features, respectively, as well as higher selection coverage in the title & country category features. The data supports the finding that the SCATTER interface helped the participants to accurately filter the attendees in the relevance-oriented task, as well as extend the selection diversity in the diversity-oriented tasks.

4.3.6 Discussion

In study of Scatter Viz, I evaluated a dual visual interface for recommending attendees at a research conference. A research conference context introduces several dimensions of attendee relevance, such as social, academic, interest, and distance similarities. Due to these factors, a traditional ensemble ranked list makes it difficult to express the diversity of recommended

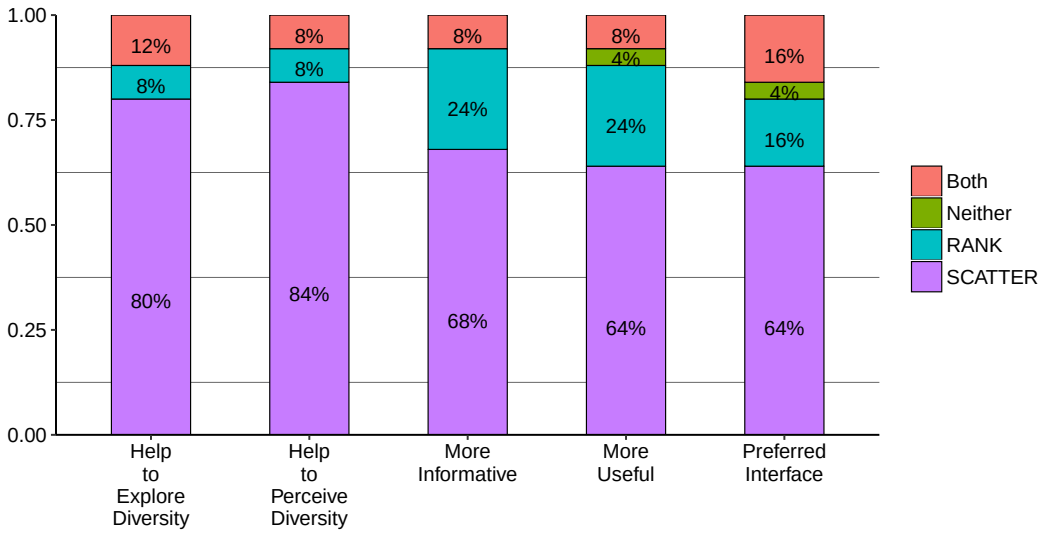


Figure 8: User preference analysis of Scatter Viz (the preferences were collected after the users experienced both interfaces). The result shows the SCATTER interface was preferred by the users in all aspects.

Table 3: Post-experiment questionnaire [60].

Q1	Which recommendation interface did you prefer?
Q2	Which recommendation interface did you find more informative?
Q3	Which recommendation interface did you find more useful?
Q4	Which recommendation interface was better at helping you to perceive the diversity of recommendations?
Q5	Which recommendation interface was better at helping you to explore the diversity of recommendation through different features and categories?

items (attendees). By spreading rankings over two dimensions, the suggested interface helps users to explore recommendations and recognize their diversity in several aspects. To assess the visual approach, I conducted a user study in a real conference environment to compare the interface (SCATTER) with a traditional ranked list (RANK) in three practical tasks. The experimental result shows a tangible incremental impact on the metrics of system usage, efficiency, and diversity.

I found that the **Scatter Viz** interface can improve user inspection on recommendations

Table 4: Diversity analysis of Scatter Viz: the table shows selection diversity for three tasks in the feature and category dimensions. The result shows that the SCATTER interface can help users to explore a more diverse set of recommendation in diversity-oriented tasks (T2 & T3). (Statistical significance level: (*) $p < 0.05$; (-) $p < 0.1$.)

	Task 1			Task 2			Task 3		
Dimensions	RANK M (SE)	SCATTER M (SE)	P	RANK M (SE)	SCATTER M (SE)	P	RANK M (SE)	SCATTER M (SE)	P
Academic + Social	0.14 (0.06)	0.11 (0.03)	*	0.14 (0.06)	0.13 (0.04)		0.14 (0.06)	0.13 (0.03)	
Academic + Interest	0.16 (0.09)	0.13 (0.08)		0.14 (0.09)	0.14 (0.08)		0.12 (0.07)	0.12 (0.08)	
Academic + Distance	0.13 (0.07)	0.12 (0.04)		0.12 (0.07)	0.14 (0.04)		0.14 (0.07)	0.16 (0.04)	
Social + Interest	0.27 (0.13)	0.21 (0.09)	*	0.25 (0.13)	0.24 (0.11)		0.22 (0.12)	0.22 (0.14)	
Social + Distance	0.27 (0.12)	0.24 (0.08)		0.26 (0.13)	0.28 (0.10)		0.24 (0.11)	0.31 (0.10)	*
Interest + Distance	0.26 (0.13)	0.25 (0.13)		0.23 (0.14)	0.27 (0.13)	-	0.21 (0.14)	0.23 (0.11)	
Title	0.17 (0.12)	0.22 (0.07)	*	0.17 (0.14)	0.31 (0.12)	*	0.17 (0.16)	0.32 (0.10)	*
Position	0.26 (0.12)	0.23 (0.10)		0.29 (0.17)	0.25 (0.14)		0.19 (0.14)	0.15 (0.08)	
Country	0.49 (0.25)	0.46 (0.15)		0.46 (0.31)	0.68 (0.26)	*	0.44 (0.26)	0.66 (0.26)	*

with multi-relevance, which leads to a higher selection diversity in the given tasks. However, I also noticed that some of the experiment participants still stick to the familiar ranked list, even when an enhanced visualization was provided. This finding helps us to realize a user preference on adopting the interface with lower learning efforts. Besides, the scatter visualization requires additional space to present, which may not be feasible in many real-world rank-based recommender systems. These findings lead to the second attempt at extending the ranked list with multi-aspect awareness, controllability, and diversity-aware designs.

4.4 CANDIDATE INTERFACE #2: RELEVANCE TUNER

The ranked list is widely applied to recommender systems for presenting recommendations to users. Even in visual recommender systems, a basic ranked list is still essential for user interactions [140, 104, 129]. A recommender system usually ranks recommended items from high to low relevance, which may reduce users' cognitive decision loading [8]. However, as mentioned above, in a context with multiple relevance aspects, a statically "ensembled" ranked list makes it more difficult for the user to recognize the impact of different aspects and to adjust the recommendations to different needs. Moreover, the persuasive design of the ranked list causes users to pay more attention to items on top of the list [26], which further decrease selection diversity and caused the fitter bubble effect though narrowing the recommendation selection [96].

As the first study shows, the users were able to properly combine relevance dimensions in a two-dimensional scatter plot visualization, but that this ability came with a steep learning curve. Despite benefits offered by the Scatter Plot, the more familiar Ranked List component was used more heavily. In the second attempt, I explored a novel interface that allowed users to explore multiple aspects of relevance within the familiar ranked list extended with a controllable tuner and stackable color bars. I proposed the **Relevance Tuner** - a visual interface with user-driven control function and meaningful visual encoding. This design expanded the ranked list representation with an ability to visualize and control multiple aspects of item relevance, which is especially important for diversity-oriented tasks. The rank-based design reduces the user's learning efforts in getting familiar with the interface.

The design is inspired by several previous research works. Ekstrand et al. [34] argued for the need to use multiple recommendation algorithms within a single system. Their approach is to let users choose the algorithm based on specific information needs. For instance, the content-based method performs better on exploring new friend while the collaborative-filtering approach is out-performed others on re-connecting old friends [17]. A social recommender system that supports multiple information needs should be controllable. In many real-world scenarios, the user might need to *fuse* multiple methods or data sources to fulfill their information needs. For example, a controllable slider has been adopted for fusing and

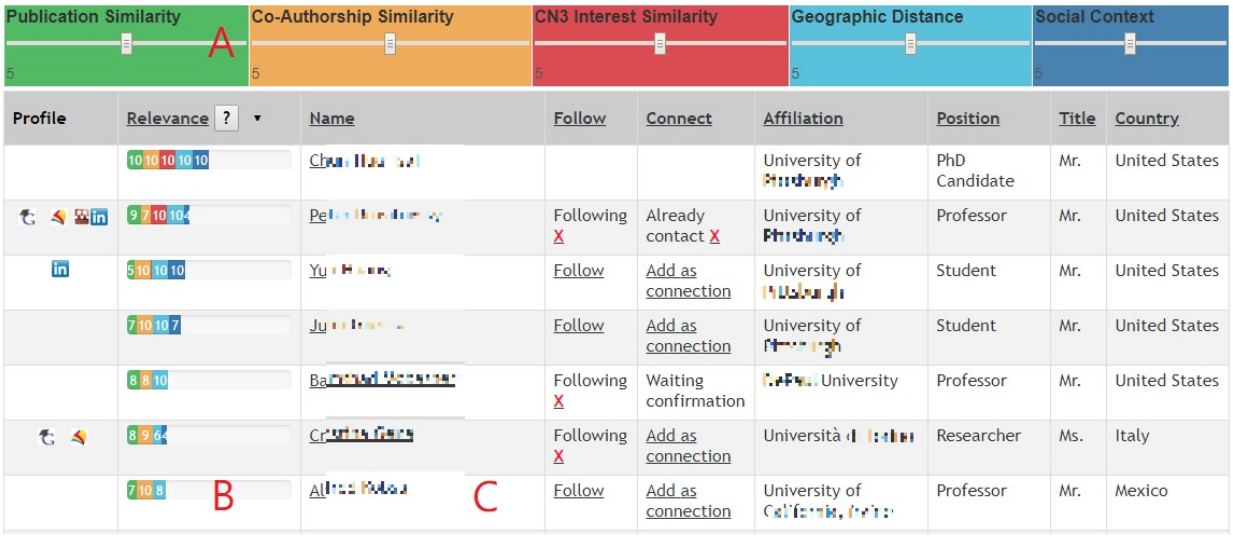


Figure 9: Prestudy: The design of Relevance Tuner: (a) Relevance Sliders; (B) Stackable Score Bar; (C) User Profiles. The interface enables the user to adjust the feature weighting on-the-fly for retrieving a customized recommendation list. The user can examine the relevant aspects of the recommended item through the multicolored score bar. (The Name and Affiliation entities have been pixelated for privacy protection.)

filtering different recommendation features, i.e., information sources, algorithms, user tags and keywords [9, 104, 142, 30]. Various kinds of stackable bars were suggested to visualize sources of algorithms that used to support a specific recommendation [104, 30, 13]. When designing **Relevance Tuner**, I intend to support source fusion with a set of relevance sliders and proportional stackable bars. The design of the Relevance Tuner is shown in Figure 9.

1. **Section A** contains five controllable sliders with the different colors representing the features of the Personalized Relevance Model. The scale of the slider ranges from 0 to 10. The user can change the weighting (W) on the fly to re-rank the ranked list below. It provides controllability for the user to adjust the ranking to different recommendation needs and preferences. The interface also adds one new feature: Social Context. This feature computes the Google search result based on the scholar's name and affiliation information; that is, the text similarity of the homepage and other related search results.

2. **Section B** shows the stackable relevance score bar for each recommended item in the ranked list. A stackable color bar interface is known for its ability to enhance controllability and transparency in a multi-aspect ranking [30]. In the system, the stackable color bars help the user to perceive how different relevance aspects of a recommended item are combined while adding transparency to the multi-aspect recommendation process. Each colored segment in the bar corresponds to one of the relevance aspects indicated by its color, which corresponds to one the sliders in **section A**. The length of the segment can vary between 0 and 20 and is determined by both, item relevance within this aspect (feature similarity) and the importance of this aspect (weight) selected by the user by adjusting sliders. The relevance score (R) is defined as:

$$R_{ij} = Round((\frac{f_{ij}}{max_j} * 10) * \frac{W_j}{5}) \tag{4.1}$$

where f_{ij} is the ith recommended item's feature similarity for the aspect j, max_j is the local maximum value of j, and W_j represents the current slider weight for j. I use the weighting percentage ($\frac{W_j}{5}$, ranged from 0 to 2) to convert the normalized feature similarity score ($\frac{f_i}{max_j} * 10$, ranged from 0 to 10) to the corresponding relevance score (R_{ij}, ranged from 0 to 20). For example, in Figure 9, the top-ranked recommendation's *Normalized Publication Similarity Score* is 10. If a user changes the *Publication Similarity* slider weight (W) to 8, then the relevance score would be $10 * \frac{8}{5} = 16$. The second recommendation's relevance score would be $9 * \frac{8}{5} = 14.4$. I will get the final relevance score as 15 after the roundup function. All relevance scores in each row will be updated, the entire list will be re-ranked by the sum of five relevance scores.

3. **Section C** shows the recommended scholar's meta-data, including name, social connection, affiliation, position, title, and country. The user can sort the ranked list by clicking the head of each column, or can inspect the explanation tabs (same as Section C in Figure 3) by clicking the name entities.

4.5 EVALUATION OF RELEVANCE TUNER

4.5.1 Data and Participants

Study of Relevance Tuner was conducted through the Conference Navigator 3 (CN3) system. The data was extended from Study of Scatter Viz to a new conference: the 25th Conference on User Modeling, Adaptation, and Personalization (UMAP 2017). A total of 20 participants (7 female) were recruited for the user study. All of the participants were attendees at the UMAP 2017 conference. They were from 15 different countries; their ages ranged from 20 to 40 (M=31.19, SE=4.97). All of them had at least one publication from UMAP 2017. The background knowledge of recommender systems score was (M=3.85, SE=0.79) on a five-point scale.

4.5.2 Experiment Design and Procedure

In Study of Relevance Tuner, I compared the interface of the ranked list plus the relevance tuner (TUNER) with a baseline of the scatter plot plus ranked list (SCATTER). The experiment design and procedure repeat the setting of Study of Scatter Viz. I manipulated the new proposed interface and adapted data from different conferences: the SCATTER interface used papers and attendees from UMAP 2017, while the TUNER interface used the same data from UMAP 2016, to minimize the learning effect between the two manipulations.

4.5.3 Action Analysis

Table 5 shows the system usage for the two interfaces of Study of Relevance Tuner. In TUNER interface, the control panel usage is defined as each time the user moves the sliders. The data supports the users interacting more frequently (it shows significance in all three tasks) with the control panel in TUNER than in SCATTER. Conversely, the users clicked more on explanation tabs in SCATTER than in TUNER. The data implies that the information listed on the table was sufficient for the users to inspect and make decisions in three proposed tasks. In task 1, the SCATTER has a significantly higher clicking frequency and

longer time spent (not significant) than the TUNER interface. The same pattern repeats in task 2 & 3, which shows that the users took more time to get familiar with the SCATTER interface. The users were gaining familiar with the TUNER interface more rapidly than with the SCATTER interface.

Table 5: User action summary of Relevance Tuner: the table shows the statistics of user interaction while solving each of the three tasks using each interface. (Statistical significance level: (*) $p < 0.05$.)

		TUNER		SCATTER		
Task	Action	M (SE)	User Count	M (SE)	User Count	
T1	Control Panel	38.4 (37.71)	20	2.85 (2.23)	18	
	Explanation Tab	9.35 (8.28)	20	22.95 (23.71)	20	*
	Click - Rank	5.05 (2.45)	20	9.8 (8.43)	17	*
	Click - Scatter	-	-	4.1 (6.03)	11	
	Time Spending	357 (289.04)	20	537 (596.98)	20	
T2	Control Panel	15.2 (13.63)	19	2 (1.71)	17	*
	Explanation Tab	6.5 (8.74)	20	8.45 (6.79)	20	
	Click - Rank	4.3 (1.21)	20	5.2 (3.76)	18	
	Click - Scatter	-	-	1.8 (2.94)	8	*
	Time Spending	201 (235.43)	20	294 (470.78)	20	
T3	Control Panel	12.2 (11.67)	17	2.25 (1.80)	19	*
	Explanation Tab	9.25 (8.75)	20	12.9 (14.38)	20	
	Click - Rank	5.15 (2.51)	20	10.2 (16.93)	17	
	Click - Scatter	-	-	2.05 (3.13)	10	*
	Time Spending	153 (92.28)	20	285 (470.78)	20	

The analysis of control and explanation function usage is reported in Figures 10 & 11 and 12. Figure 10 shows the re-weighting frequency of the TUNER interface. The red line is the usage of Tuner sliders, and the blue line is the average feature score selected by the user during re-tuning. The data indicates that the user tends to interact with the sliders

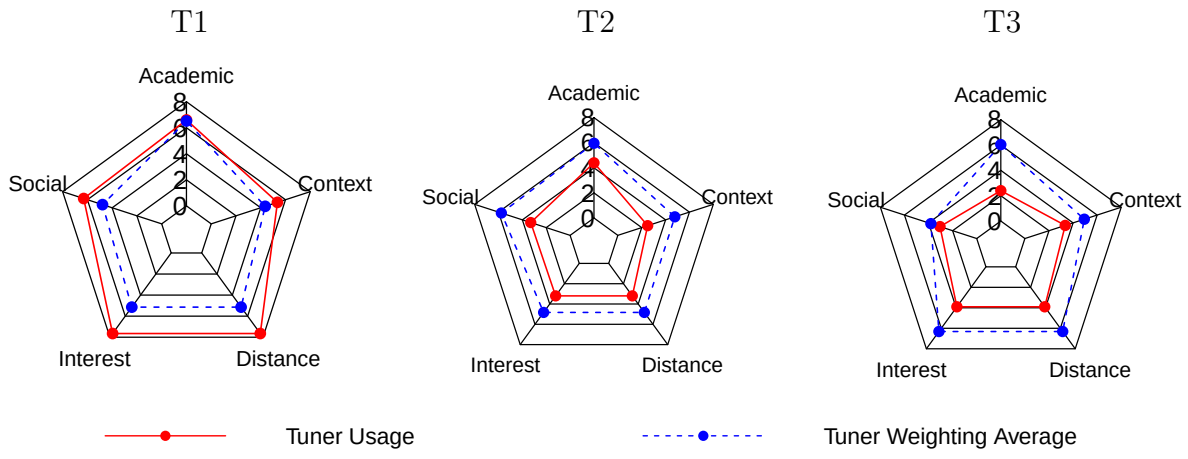


Figure 10: Relevance slider usage of Relevance Tuner: The red line is the tuner usage, while the blue line is the average weighting score of the overall re-tuning.

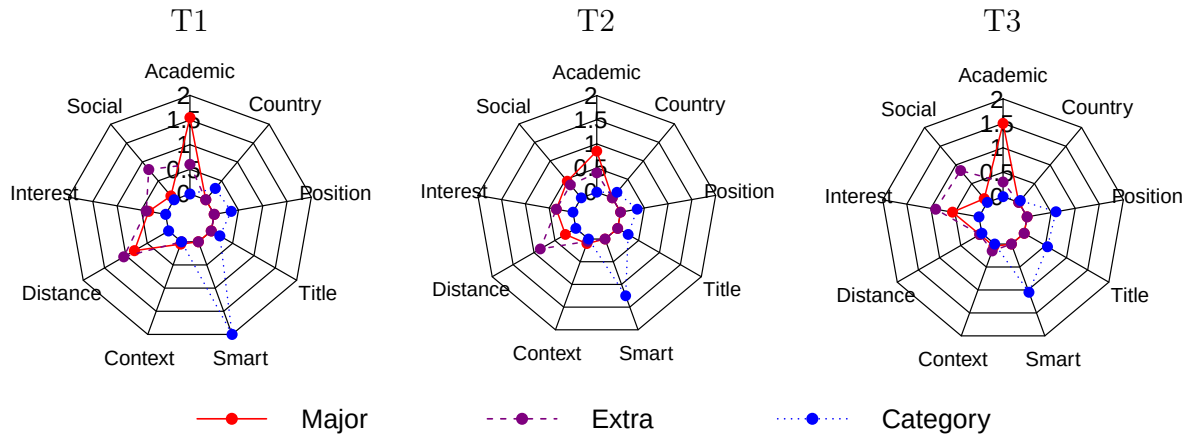


Figure 11: Scatter plot features usage of Relevance Tuner: The red line is the usage of major features, the purple line is the usage of extra features, and the blue line is the usage of category features.

more in the beginning (task1), but interacts less in a later task (2 & 3). However, the overall amount of manipulation with features was remarkably higher for TUNER than SCATTER (Figure 11) for all tasks. The average weighting scores for TUNER show only a very slight

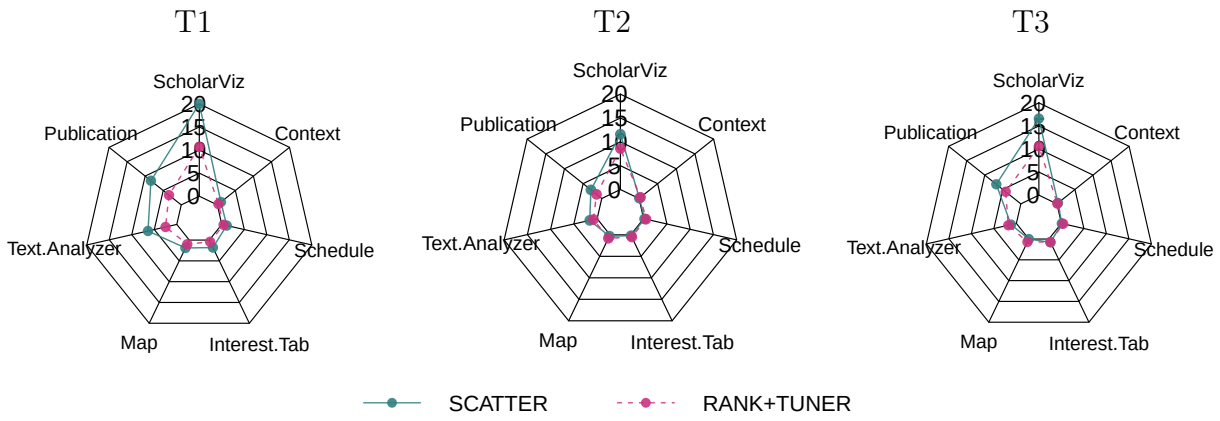


Figure 12: Explanation function usage of Relevance Tuner: The light plum color is the usage of RANK interface, while the light blue color is the usage of SCATTER interface.

association with the nature of the task at hand; that is, all features were potentially useful to filter the recommendation result in both relevance and diversity-oriented tasks. In contrast, the relatively rare use of SCATTER features (Figure 11) showed a consistent pattern with Study of Scatter Viz - the users selected the recommendation features based on the tasks' requirement. The comparative use of explanations shown in Figure 12 is also consistent with Study of Scatter Viz showing that the main demand was for *socialviz* and *publication* tabs. At the same time, it is interesting to observe that users in the TUNER group requested explanation less frequently than the SCATTER group, as through a regular TUNER interface provided more information for decision making than the SCATTER interface. However, I also observed that the TUNER users took less time to finish the tasks, as compared to those using the SCATTER interface.

4.5.4 User Feedback Analysis

Figure 13 shows the analysis of the post-stage survey. The high rating in both interfaces shows the positive user acceptance in Study of Relevance Tuner (no significance on all the factors). However, the user tends to favor the TUNER interface when considering the factors

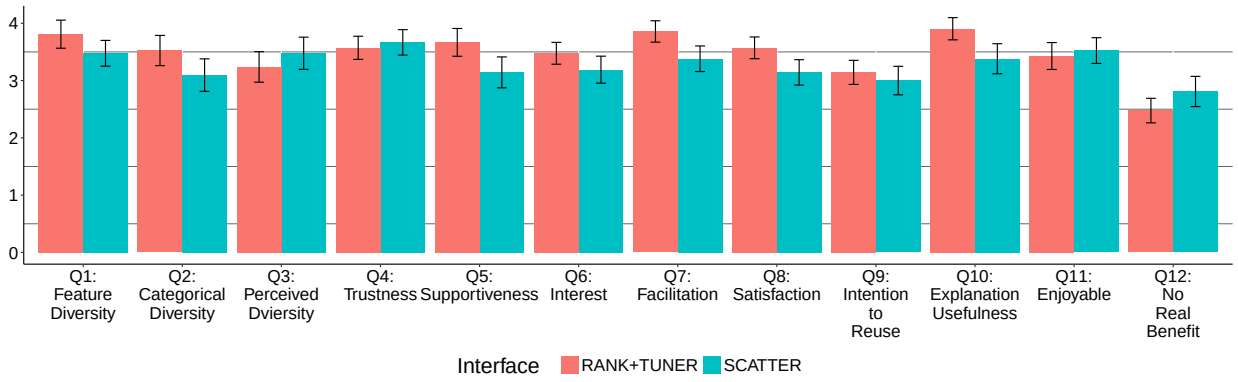


Figure 13: User feedback analysis of Relevance Tuner: I did not find significant difference in all aspects, which indicates that the usability of two interfaces was comparable. (A cut-off value was set at 3.5 on the 5 point scale. Statistical significance level: (*) $p < 0.05$.)

of Supportiveness (Q5), Interest (Q6), Facilitation (Q7), Satisfaction (Q8), Intend to Reuse (Q9), and Usefulness (Q10). The SCATTER interface performs better on the measures of Trustiness (Q4) and Enjoyable (Q11). This result supports the users in favor of rank-based list more than the visual-based interface, but the visualization shows an increased level of usability on gaining trust and enjoyment in using the interface. Surprisingly, the feedback also indicates that the TUNER interface would be better for the user to fulfill the task on the feature diversity (Q1) and category diversity (Q2), but that the SCATTER interface is outperformed on the ability to Perceive Diversity (Q3).

This result shows that a user tends to use the ranked list with better controllability and transparency to conduct diversity-oriented tasks. The scatter visualization would play the role of helping the users to perceive diversity in multiple areas of relevance. The final preference result in Figure 14 also confirms this conclusion. About half of the users select the TUNER interface as the one with an advantage at helping to explore diversity, providing more informative information, being more useful, and fitting their preference - but the users also agree that the SCATTER interface could better help to perceive diversity after they finished the three tasks on two interfaces.

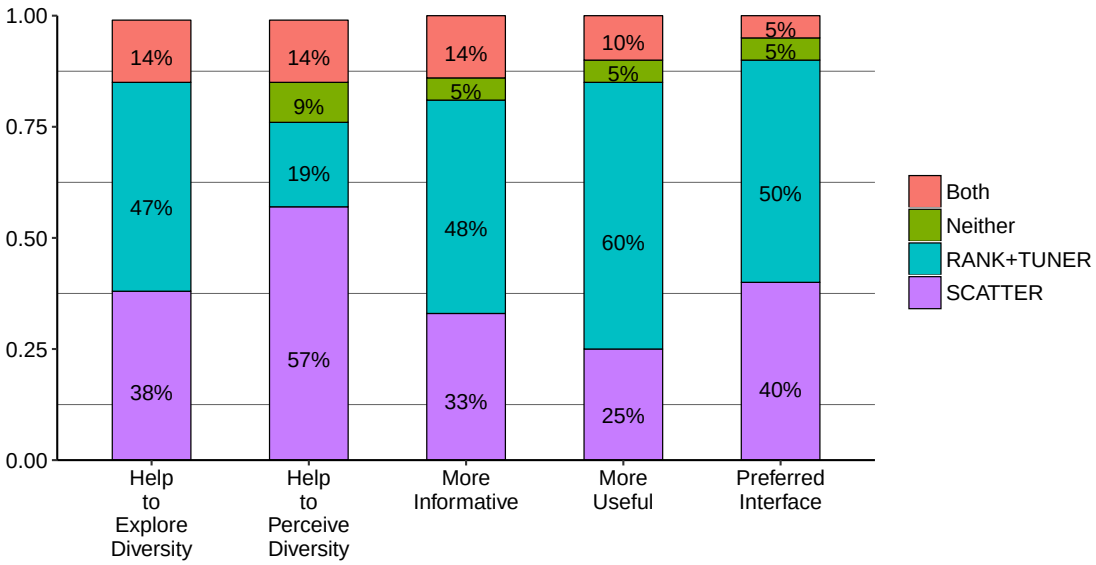


Figure 14: User preferences analysis of Relevance Tuner collected after the users experienced both interfaces. The result shows that the TUNER interface was preferred by users in all aspects except perceived diversity.

4.5.5 Recommendation Diversity Analysis

Table 6 shows the diversity analysis of Study of Relevance Tuner. In task 1 (relevance-oriented), there is no significant difference between the two interfaces on the diversity measurement, showing both of the interfaces can support the user to fulfill a relevance-oriented task. However, in the diversity-oriented tasks 2&3, I found the TUNER group could achieve higher entropy than the SCATTER group. This finding hints that even with a ranked-list interface the user can achieve a reasonable level of selection diversity if controllability and transparency for each considered dimension of relevance are available. At the same time, the SCATTER interface performed slightly (but not significantly) better than the TUNER interface in the *category* diversity metrics. This finding helps to highlight the value of color-coding data in the SCATTER interface, a feature not supported by TUNER. The diversity analysis of Social Context is omitted due to the page limitation.

Table 6: Diversity analysis of Relevance Tuner: the table shows selection diversity for three tasks for each feature combination and category dimensions. The result indicates that the TUNER interface enabled users to explore a more diverse set of recommendation in diversity-oriented tasks T2 & T3. (Statistical significance level: (*) $p < 0.05$.)

	Task 1			Task 2			Task 3		
Dimensions	TUNER M (SE)	SCATTER M (SE)	P	TUNER M (SE)	SCATTER M (SE)	P	TUNER M (SE)	SCATTER M (SE)	P
Academic + Social	0.12 (0.05)	0.12 (0.05)		0.15 (0.02)	0.12 (0.04)	*	0.16 (0.03)	0.13 (0.05)	
Academic + Interest	0.17 (0.10)	0.16 (0.07)		0.17 (0.07)	0.13 (0.07)	*	0.14 (0.07)	0.14 (0.06)	
Academic + Distance	0.13 (0.06)	0.12 (0.04)		0.16 (0.06)	0.12 (0.05)	*	0.17 (0.05)	0.14 (0.05)	*
Social + Interest	0.23 (0.11)	0.23 (0.10)		0.25 (0.07)	0.19 (0.08)	*	0.32 (0.09)	0.26 (0.16)	
Social + Distance	0.23 (0.10)	0.22 (0.09)		0.28 (0.07)	0.22 (0.10)	*	0.32 (0.08)	0.26 (0.09)	*
Interest + Distance	0.24 (0.16)	0.29 (0.12)		0.31 (0.15)	0.24 (0.13)	*	0.28 (0.12)	0.26 (0.13)	
Title	0.12 (0.04)	0.16 (0.13)		0.15 (0.02)	0.17 (0.12)		0.15 (0.02)	0.14 (0.05)	
Position	0.29 (0.15)	0.27 (0.12)		0.28 (0.12)	0.32 (0.15)		0.16 (0.07)	0.22 (0.16)	
Country	0.35 (0.20)	0.57 (0.29)	*	0.41 (0.24)	0.58 (0.22)	*	0.48 (0.19)	0.54 (0.32)	

4.5.6 Discussion

In Study of Relevance Tuner, I presented a new rank-based interface for recommender attendees at a research conference. A total of five dimensions of relevance were proposed from the Personalized Relevance Model. I conducted a user study in a real conference environment to compare the two interfaces of an enhanced ranked list (TUNER) and the visualization interface (SCATTER). The experimental results suggested the different suitable scenarios for the two interfaces. I found that, even in diversity tasks with multi-relevance settings, the users were still able to fulfill the diversity task with a rank-based interface, but it required the support of interface controllability and transparency through visual encoding. Besides, while I found that the user would better perceive the diversity in the SCATTER interface, the user would prefer to adopt the TUNER interface to fulfill the diversity tasks.

Furthermore, when the user was interacting with the TUNER interface, the user spends more time inspecting the information on each row, instead of checking the explanation functions. This result shows that the higher level of diversity exploration was not triggered by the diversity-enhanced visualization or explanation (fewer clicks on the explanation tabs), but was instead contributed by the intention of the user reaction to the simulated diversity-oriented tasks. In the SCATTER interface, the user relies more on the explanation function and multi-relevance visualization to explore diversity-oriented tasks. Although the result showed lower entropy measurement when the user adapted to the SCATTER interface, the SCATTER interface can better help the user to perceive the diversity among multiple-relevance dimensions, based on the user feedback analysis through post-study questionnaires.

4.6 SUMMARY

In this chapter, I presented experiments with three different social recommender interfaces: Ranked List (RANK), *Scatter Viz* (SCATTER) and *Relevance Tuner* (TUNER). I first showed that providing a scatter plot (SCATTER) can help the user to better fulfill the diversity-related tasks, as compared to a simple ranked list (RANK). However, despite the benefits of the new two-dimensional presentation, the users still extensively used the ranked list component of the interface.

Based on the results of the Scatter Viz, I attempted to integrate the ability to coordinate multiple aspects of relevance within the ranked list rather than offering it in a separate component as in SCATTER. To compensate for the biasing nature of the ranked list, I also provided a controllable fusion of relevance aspects. The resulting expanded ranked list interface (TUNER) offered both controllability and visual encoding of multiple relevance aspects. I showed that the users could adopt a rank-based list to fulfill diversity-oriented tasks with higher selection diversity. The usability analysis revealed that both SCATTER and TUNER were ranked by the conference users with high subjective ratings. However,

TUNER required less learning effort. I also discussed the mediation effects of the proposed interfaces on the user experience. The analysis helps to describe the benefits of the two proposed interfaces in social recommender systems.

One of the goals was to understand how users apply the recommender interfaces to diversity-oriented tasks. I prepared tasks that encourage users to explore the conference attendees using multiple aspects of relevance. I found that both of the proposed interfaces were capable of helping the user to fulfill the assigned tasks. The experimental result supported the finding that the participants were able to correlate multiple aspects of relevance using two dimensions of visualization in SCATTER and the controllable ranked list with multi-aspect visualization in TUNER. I found that an extension of a traditional ranked list with controllability and visualization was better than a separate visualization component in the sense of on getting familiar with the new interface (which led to a higher rating on the user preference). In contrast, a separate diversity-enhanced visualization can also achieve the goal, but it came at the cost of a steeper learning curve. However, once the users were familiar with the interface, it brought an advantage of helping the users to perceive diversity and gain trust in the recommendations.

The study has some limitations. First, the within-subject user study was conducted using consecutive years of the same conference series. Some well-known and senior domain experts may appear in the recommendation list for both conferences. This repetition may cause bias in the user studies. Second, the data sparsity and cold-start problem may hurt the recommendation performance; for example, the Interest feature is less useful for users who never bookmarked any talks or papers within the Conference Navigator System (CN3). I tried my best to send out emails both before and during the conference to improve interest-based recommendations. Third, the scale of reported user studies is relatively small. It may decrease the statistical power of the findings. Fourth, the experiment was conducted at mid-size conferences, so I was not able to explore scaling issues which might occur at conferences with a much larger number of attendees or in a different recommendation context with a large set of items to explore.

5.0 DESIGNING EXPLANATION INTERFACES USING STAGE-BASED PARTICIPATORY DESIGN APPROACH

In this chapter, I presented the first three stages of a stage-based participatory process [32] for designing explanation interfaces in a hybrid social recommender system. I report the findings of *Expert Mental Model*, *User Mental Model* and *Target User Model* that determined *1) what should be explained of the recommendation models*, *2) current user expectation* and *3) new user expectation* of the research platform system, respectively. The finding of this chapter is used to design the prototype interfaces and the card-sorting factors in Chapter 6.

5.1 INTRODUCTION

Instead of the offline performance improvements, more and more researches focused on the works of evaluating the system from the *user experience*, i.e., what is the user perception on the explanation interfaces? Explaining recommendations (i.e., enhancing the system explainability) can achieve different *explanatory goals*, which help users to make a better decision or persuading them to accept the suggestions from a system [125, 116]. I followed the seven explanatory goals that proposed by [126]: *Transparency*, *Scrutability*, *Trust*, *Persuasiveness*, *Effectiveness*, *Efficiency*, and *Satisfaction*. Since it is hard to have a single explanation interface that achieves all these goals equally well, the designer needs to make a trade-off while choosing or designing the form of interface [126]. For instance, an interactive interface can be adapted to increase user trust and satisfaction but may prolong the decision and explore process while using the system (i.e., lead to decreasing of efficiency) [132].

Over the past few years, several approaches have been discussed to enhance the explain-

ability in the recommender systems. The approaches can be summarized by different styles, reasoning models, paradigms, and information [37]. 1) *Styles:* [75] conducted an online user survey to explore the user preference in nine explanation styles. They found *Venn diagrams* outperformed all other visual and text-based interfaces. 2) *Reasoning Models:* [141] used tags to explain the recommended item and the user's profile. The approach emphasized the factor of why a specific recommendation is plausible, instead of revealing the process of recommendation or data. *3) Paradigms:* [58] presented a model for explanations based on the user's conceptual model of the collaborative-based recommendation process. The result of the evaluation indicates two interfaces - "Histogram with grouping" and "Presenting past performance" - improved the acceptance of recommendations. 4) *Information:* [107] proposed explanations tailored to the user and recommendation, i.e., although one recommendation is not the most popular one, the explanation would justify the recommendation by providing the reasons.

Although many approaches have been proposed to enhance the recommender explainability, bringing explanation interfaces to an existing recommender system is still a challenging task. More recently [32] suggested a different approach to improve *user mental model (UMM)* while bringing transparency (explanations) to a recommender system. The model described the process of a user builds an internal conceptualization of the system or interface along with user-system interactions, i.e., building the knowledge of how to interact with the system. If the model is misguided or opaque, the users will face difficulties in predicting or interpreting the system [32]. Hence, the researchers suggested to improve the *mental model*, so the users can gain *awareness* while using the system as well as the explanation interfaces.

I adopted the stage-based participatory framework from [32], which intends to answer two key questions while designing the explainable user interface (UI): a) What to Explain? And b) How to explain?

The process can be summarized in four stages.

1. *Expert Mental Model:* What can be explained? I defined an expert as the recommender system developer.
2. *User Mental Model:* What is the user mental model of the system based on its current UI? The model should be built through the current recommender system users.

3. *Target Mental Model:* Which key components of the algorithm do users want to be made explainable in the UI? The *target user* is the users who are new to the system.
4. *Iterative Prototyping:* How can the target mental model be reached through UI design. The key is to measure if the proposed explanation interfaces achieved the explanatory goals.

5.2 FIRST STAGE: EXPERT MENTAL MODEL

I adopted a hybrid explanation approach [103, 75], which mixed multiple visualizations to explain the details of the recommendation model. I want to let the users understand both a) the mutual relationship (similarity) between him/herself and the recommended scholar and b) the key component in each recommendation model. I then discussed the *Expert Mental Model* through the system developing process of the five recommendation models.

1. **Publication Similarity:** The similarity was determined by the degree of text similarity between two scholars' publications using cosine similarity. I applied tf-idf to create the vector with a word frequency upper bound of 0.5 and a lower bound of 0.01 to eliminate both common and rarely used words. In this model, the key components were the *terms* of the paper title and abstract as well as its *term frequency*.
2. **Topic Similarity:** This similarity was determined by matching research interests using topic modeling. I used latent Dirichlet allocation (LDA) to attribute collected terms from publications to one of the topics. I chose 30 topics to build the topic model for all scholars. Based on the model, I then calculated the topic similarity between any two scholars. The key components were the *research topics* and the *topical words* of each research topic [148].
3. **Co-Authorship Similarity:** This similarity approximated the network distance between the source and recommended users. For each pair of the scholar, I tried to find six possible paths for connecting them, based on their coauthorship relationships. The network distance is determined by the average distance of the six paths. The key components

were the *coauthors* (as nodes), *coauthorship* (as edges) and the *distance of connection the two scholars*.

4. **CN3 Interest Similarity:** This similarity was determined by the number of co-bookmarked conference papers and co-connected authors in the experimental social system (CN3). I simply used the number of shared items as the CN3 interest similarity. The key component is the shared *conference papers* and authors.
5. **Geographic Distance:** This similarity was a measurement of the geographic distance between attendees. I retrieved longitude and latitude data based on attendees' affiliation information. I used the Haversine formula to compute the geographic distance between scholars. The key components are the *geographic distance* and *affiliation information of the scholars*.

5.3 SECOND STAGE: USER MENTAL MODEL

As a first step towards understanding the design factors of explanatory interfaces, I deployed a survey through a social recommender system, Conference Navigator [131], and analyzed data from the respondents. I targeted the users who had created an account and interacted with the system in their previous conference attendance (at least using the system for one conference). The survey was initiated by sending an invitation to the qualified users in December 2017. I sent out 89 letters to the conference attendees of UMAP/HT 2016, and a total of 14 participants (7 female) replied to create the pool of participants for the user study. The participants were from 13 different countries; their ages ranged from 20 to 40 (M=31.36, SE=5.04). I did an online survey to collect necessary demographic information and self-reflection about how to design an explanation function in seven explanatory goals [126].

The proposed questions were:

How can an explanation function help you to perceive system ...

1. **Transparency** - explain how the system works?
2. **Scrutability** - allow you to tell the system it is wrong?

3. **Trust** - increase your confidence in the system?
4. **Persuasiveness** - convince you to explore or to follow new friends?
5. **Effectiveness** - help you make good decisions?
6. **Efficiency** - help you to make decisions faster?
7. **Satisfaction** - make using the system fun and useful?

I asked the participants to answer each question in 50-100 words, in particular, reflecting the explanatory goals of the social recommendation. The data was published in [133].

1) Transparency: 71% of respondents pointed out the *reasons* of generated social recommendations that help them to perceive higher system transparency, i.e., the personalized explanation, the linkage and data sources, reasoning method and understandability. I then summarized the feedback into five factors: *1) The visualization presents the similarity between my interest and the recommended person. 2) The visualization presents the relationship between the recommended person and me. 3) The visualization presents where did the data were retrieved. 4) The visualization presents more in-depth information on how the score amounts up. 5) The visualization allows me to see the connections between people and understand how they are connected.*

2) Scrutability: Half of the respondents mentioned they needed "inspectable details" to figure out the wrong recommendation. 35% of respondents suggested the mechanism of accepting user feedback on improving wrong recommendations, such as a space to submit user ratings or yes/no options. 14% of respondents preferred a dynamic exploration process to determine the recommendation quality. I then summarized the feedback into four factors: *6) The visualization allows me to understand whether the recommendation is good or not. 7) The visualization presents the data for making the recommendations. 8) The visualization allows me to compare and decide whether the system is correct or wrong. 9) The visualization allows me to explore and then determine the recommendation quality.*

3) Trust: 28% of respondents mentioned that they trusted the system more when they perceived the benefits of using the system. 35% of respondents preferred to trust a system with reliable and informative explanations, more detailed information or understandable. 35% of respondents mentioned they trust a system with transparency or passed their verification. I then summarized the feedback into three factors: *10) The visualization presents*

a convincing explanation to justify the recommendation. 11) The visualization presents the components (e.g., algorithm) that influenced the recommendation. 5) The visualization allows me to see the connections between people and understand how they are connected.

4) Persuasiveness: Half of the respondents mentioned the explanation of social familiarity would persuade them to explore novel social connections, namely, when shown social context details or shared interests. 21% of respondents indicated that an informative interface could boost the exploration of a new friendship. 28% of respondents preferred a design that inspired curiosity, implicit relationships. I then summarized the feedback into three factors: *12) The visualization shows me the shared interests, i.e., why my interests are aligned with the recommended person. 13) The visualization has a friendly, easy-to-use interface. 14) The visualization inspired my curiosity (to discover more information).*

5) Effectiveness: 64% of respondents mentioned that the aspects of social recommendation relevance helped them to make a good decision. The aspect included explaining the recommendation process, understandable or more informative. 28% of respondents suggested a reminder that a historical or successful decision could help them to make a good decision, i.e., a previously-made user decision and success stories. I then summarized the feedback into three factors: *15) The visualization presents the recommendation process. 5) The visualization allows me to see the connections between people and understand how they are connected. 11) The visualization presents the components (e.g., algorithm) that influenced the recommendation.*

6) Efficiency: 28% of respondents mentioned that a proper highlighting of the recommendation helped to make the decision faster. For example, they are emphasizing the relatedness, identifying the top recommendations or providing success stories. 28% of respondents preferred a tune-able or visualized interface to accelerate the decision process, such as tuning the recommendation features, visualizing the recommendations. However, the explanations may not always be useful. 21% of respondents argued that the explanation would prolong the decision process instead of speeding it up: the user may need to take extra time to examine the explanations. I then summarized the feedback into two factors: *16) The visualization presents highlighted items/information that is strongly related to me. 17) The visualization presents aggregated, non-obvious relations to me.*

7) Satisfaction: The feedback on how an explanation can help the user satisfy the system was varied. Three aspects received an equal 7% of respondents' preferences. That is, users preferred to view the feedback from the community, shown the historical interaction record, and provided a personalized explanation. Two aspects received an equal 14% of respondents' preference, i.e., a focus on a friendly user interface and saved decision time. 21% of respondents reported a higher satisfaction on using the explanation as a "small talk topic", i.e., as an initial conversation in a conference. 28% of respondents preferred an interactive interface for perceiving the system to be fun, e.g., a controllable interface. I then summarized the feedback into four factors: *18) The visualization presents the feedback from other users, i.e., I can see how others rated the recommended person. 19) The visualization allows me to tell why does this system recommends the person to me. 1) The visualization presents the similarity between my interest and the recommended person. 13) The visualization is a friendly, easy-to-use interface.*

Based on the result of the online survey, I concluded a total of 19 factors in the second stage of building the user mental model.

1. The visualization presents the similarity between my interest and the recommended person.
2. The visualization presents the relationship between the recommended person and me.
3. The visualization presents where the data was retrieved.
4. The visualization presents more in-depth information on how the scores sum up.
5. The visualization allows me to see the connections between people and understand how they are connected.
6. The visualization allows me to understand whether the recommendation is good or not.
7. The visualization presents the data for making the recommendations.
8. The visualization allows me to compare and decide whether the system is correct or wrong.
9. The visualization allows me to explore and then determine the recommendation quality.
10. The visualization presents a convincing explanation to justify the recommendation.
11. The visualization presents the components (e.g., algorithm) that influenced the recommendation.

12. The visualization shows me the shared interests, i.e., why my interests are aligned with the recommended person.
13. The visualization has a friendly, easy-to-use interface
14. The visualization inspired my curiosity (to discover more information).
15. The visualization presents the recommendation process clearly.
16. The visualization presents highlighted items/information that is strongly related to me.
17. The visualization presents aggregated, non-obvious relations to me.
18. The visualization presents feedback from other users, i.e., I can see how others rated a recommended person.
19. The visualization allows me to tell why does this system recommends the person to me.

I also found some factors across different exploratory goals. For example, Factor 1 were shared by the exploratory goal of *Transparency* and *Satisfaction*. Factor 5 were shared by *Transparency*, *Trust* and *Effectiveness*. Factor 11 was shared by *Trust* and *Effectiveness*. Factor 13 was shared by *Persuasiveness* and *Satisfaction*.

5.4 THIRD STAGE: TARGET MENTAL MODEL (STUDY 1)

In this stage, I conducted a controlled lab study 1a for creating the *Target Mental Model*. The model is used to identify the key components of the recommendation model that the users might want to be explainable in the UI. Since the goal is to identify the information need for new users, I specifically selected subjects who never used the CN3 system. A total of 15 (6 female) participants (N=15) were recruited for this study. They are first, or second-year graduate students (major in information sciences) at the University of Pittsburgh with ages ranged from 20 to 30 (M=25.73, SE=2.89). All participants had no previous experience of using the CN system. Each participant received USD$20 compensation and signed an informed consent form.

I asked the subjects to complete a card-sorting task about their preference for the 19 factors I identified in the second stage. I started by presenting the CN3 system (shown in Figure 9) to the subjects and introducing the five recommendation models through the

Expert Mental Model. After the tutorial, the subjects were asked to do a closed card-sorting that assigns cards into four predefined groups. The four groups are 1) very important; 2) less important; 3) not important, and 4) not relevant.

The survey result is reported in Table 7. I found that for the target users, factors 1, 13, 16 outperformed other factors: more than ten subjects assigned the three factors into the "very important" group. The factor 2, 6, 10, 12, 14, 15 and 19 formed the secondary preference group with at least 10 subject assigning them into "very important" or "less important" groups. The subject's least preferred factor were 3, 7, 11, 18 with at least nine subjects assigning these factors into "not important" or "not relevant" groups.

Based on the card-sorting result, I found the user preferred an explainable UI is presenting the similarity between his/her interests and the recommended person (F1). The UI should be friendly and easy-to-use (F13) as well as highlighted the items or information that is strongly related to the user (F16). Besides, some factors are also liked by the subjects. For instance, the UI is presenting the mutual relationship (F2), shared interests (F12), and recommendation process (F15). The UI should also allow the user to understand (F6) and justify (F10) the quality of recommendation as well as inspired the curiosity of exploration (F14) and recommendation process (F19). Interestingly, I also found the users were less interested in a UI of presenting the data source (F3) and raw data (F7) as well as the detail of algorithm (F11) and the recommendation feedback from the other users in the same community (F18).

Hence, I decide to filter out the factors that were less preferred by the subjects. I choose to keep the factors with more than ten votes in the groups of "Very Important" and "Less Important", which are F1, F2, F6, F10, F12, F13, F14, F15, F16, F19, the chosen factors were highlighted in red color in Table 7. The factors can be projected back to the original explanatory goals. The mentioned percentage of each exploration goal is listed as below: Transparency (40%, 2 out of 5), Scrutability (0%, 0 out of 4), Trust (33%, 1 out of 3), Persuasiveness (67%, 2 out of 3), Effectiveness (33%, 1 out of 3), Efficiency (50%, 1 out of 2) and Satisfaction, (75%, 3 out of 4). That is, the *Target Mental model* was built through the exploratory goal of (rank from high to low importance) Satisfaction, Persuasiveness, Efficiency, Transparency, Trust, and Effectiveness.

5.5 SUMMARY

In this chapter, I presented a participatory process of bringing explanation interfaces to a social recommender system. I proposed three stages in responding to the challenge questions in identifying the key components of explanation models and mental models. In the first stage, I discussed the *Expert Mental Model* by discussing the key components (based on the similarity algorithm) of each recommendation model. In the second stage, I reported an online survey of current system users (N=14) and identified 19 explanatory goals as the *User Mental Model*. In the third stage, I reported the card-sorting results of a controlled user study (N=15) that created the *Target Mental Model* through the target users' preference of the explanatory factors.

Table 7: The target users card-sorting results

	Very Important	Less Important	Not Important	Not Relevant
Factor 1	11	1	3	0
Factor 2	9	5	1	0
Factor 3	0	2	10	3
Factor 4	1	8	3	3
Factor 5	5	4	6	0
Factor 6	7	6	2	0
Factor 7	3	2	9	1
Factor 8	4	3	3	5
Factor 9	7	2	4	2
Factor 10	3	9	2	1
Factor 11	0	6	6	3
Factor 12	4	6	5	0
Factor 13	13	2	0	0
Factor 14	0	13	2	0
Factor 15	4	7	3	1
Factor 16	10	5	0	0
Factor 17	3	6	3	3
Factor 18	1	5	5	4
Factor 19	1	10	3	1

6.0 EVALUATING PROTOTYPES OF EXPLANATION INTERFACES

This chapter presents my two stages of the investigation of iteratively implement and evaluate explanation interfaces for five recommendation models. In this first stage, I introduce a total of 25 prototype interfaces for the five recommendation models and report the card-sorting result of study 2. In the second stage, I conduct the first round of evaluation identified the effective design for the five recommendation models through study 3.

6.1 FIRST STAGE: ITERATIVE PROTOTYPING (STUDY 2)

After the card-sorting task of study 1, I asked the same group of subjects to identify the chosen ten factors across some UI prototypes, as study 2. A total of 15 (6 female) participants (N=15) were recruited. They were first, or second-year information science graduate students at the University of Pittsburgh with ages ranged from 20 to 30 (M=25.73, SE=2.89). All participants had no previous experience of using the CN system. Each participant received USD$20 compensation and signed an informed consent form. Subjects took between 40 and 60 minutes to complete the study.

A total of 25 interfaces (five interfaces for each recommendation model) were exposed in this stage. I used a within-subject design, i.e., all participants required to do a card-sorting task. In each session, the participants were asked to sort the given five interfaces into groups 1 to 5 (1: Strongly Agree, 5: Strongly Disagree), in each exploratory factor. If one interface is not contributing to the factor, the participant can mark it as irrelevant (not applicable). I continued with a semi-interview after the subject completed each session to collect the qualitative feedback.

I conducted study 2 to determine the *user preferred* visual interfaces of explaining the five similarity-based recommendation models (E1, E2, E3, E4 & E5). The participants were asked to complete *closed card-sorting tasks* to organize the proposed interfaces into predefined groups. The tasks were designed to evaluate *how well a visual interface supports the exploratory goal*. A total of nineteen factor across seven explanatory goals were introduced in the study [125, 133]. The seven factors included Transparency (**TP**), Scrutability (**SC**), Trust (**TS**), Persuasiveness (**PE**), Effectiveness (**ET**), Efficiency (**EF**) and Satisfaction (**SA**). The detailed statement of each factor can be found in this section.

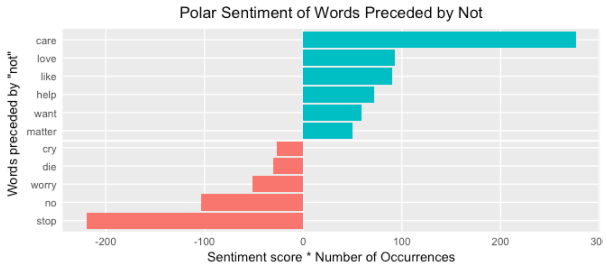
There were a total of five card-sorting sessions for all five recommendation model. At the beginning of the study, I introduced the CN system and the recommendation models (from *Expert Mental Model*) to the subjects. After the introduction, I asked the subjects to complete a closed card-sorting task for each recommendation model. In each task, I presented five explanation interfaces (paper mock-ups) and asked the subjects to assign the interfaces to group 1-5 (from Group 1: Strongly Agree; to Group 5: Strongly Disagree, or Not Applicable) based on the given exploratory factors (listed in Table 9). The experiment followed the within-subject design, i.e., all participants required to perform three card-sorting tasks (i.e., one for each group) with the same nineteen explanatory factors. The order of tasks and factors was the same for all participants. I continued with a semi-interview after each task to collect the qualitative feedback.

6.1.1 Explaining Publication Similarity

The key component of publication similarity is *terms* and *term frequency* of the publication as well as its mutual relationship (i.e., the common terms) between two scholars. I presented four visual interface prototypes (shown in Figure 15) for explaining publication similarity and one text-based interface (E1-1), which simply says "You and [the scholar] have common words in [W1], [W2], [W3]."

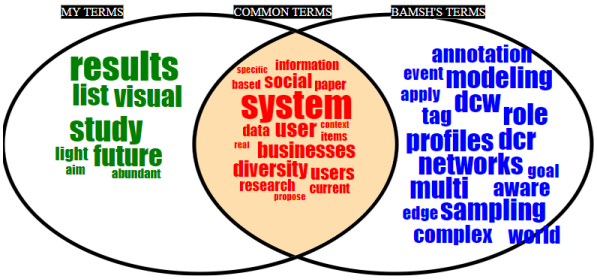
E1-2: Two-way Bar Chart The bar chart is a common approach in analyzing the text mining outcome [118] using a histogram of terms and term frequency. I extended the design to a two-way bar chart to show the mutual relationship of two scholars' publication

(a) E1-2: Two-way Bar Chart



(b) E1-3: Word Clouds

(c) E1-4: Venn Word Cloud



(d) E1-5: Interactive Word Cloud

Figure 15: The prototype interfaces for *Publication Similarity* in study 2.

terms and term frequency, i.e., one scholar in positive and the other scholar on a negative scale. The design is shown in Figure 15a.

E1-3: Word Clouds Word cloud is a common design in explaining text similarity [131]. I adopted the word cloud design from [149], which presented the term in the cloud and the term frequency by the font size. This interface provided two word clouds (one for each scholar) so the user can perceive the mutual relationship. The design is shown in Figure 15b.

E1-4: Venn Word Cloud Venn diagram was recognized as an effective hybrid explanation interface by [75]. This interface could be considered as a combination of a word cloud and a Venn diagram [136], which presents term frequency using the font size. The unique terms of each scholar are shown in a different color (green and blue) while the common terms are presented in the middle, with red color, for determining the mutual relationship. The

design is shown in Figure 15c.

E1-5: Interactive Word Cloud A word cloud can be interactive. I extend the idea from [131] and used Zoomdata Wordcloud [154], which follows the common approach to visualize term frequency with the font size. The font color was selected to distinguish the scholars' terms, i.e., different term colors for each scholar. A slider was attached to the bottom of the interface that provides real-time interactive functionality to increase or decrease the number of terms in the word cloud. The design was shown in Figure 15d.

Results The card-sorting result was presented in Table 8. I found the *E1-4 Venn Word Cloud* was preferred by the participants, received 76 votes in Rank 1, which was outperformed the other four interfaces. According to the post-session interview, 13 subjects agreed E1-4 is the best interface versus the other four interfaces. The supporting reasons can be summarized as 1) the Venn diagram provided common terms in the middle, which highlighted the common terms and shared relationship; 2) it is useful to show non-overlapping terms on the sides (N=5) and 3) the design is simple, easy to understand and require less time to process (N=3). Two subjects mentioned they preferred E1-2 the most due to histograms gives them the "concrete numbers" for "calculating" the similarity, which was harder when using word clouds.

6.1.2 Explaining Topic Similarity

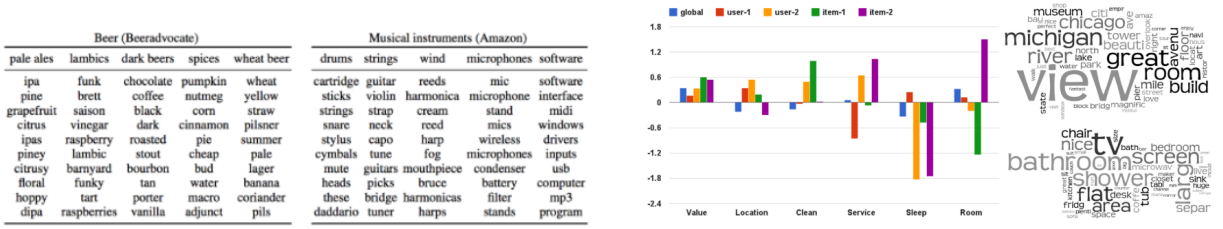
The key component of topic similarity is *research topics* and *topical words* of the scholar as well as its mutual relationship (i.e., the common research topics) between two scholars. I presented four visual interfaces prototypes (shown in Figure 16) and one text-based prototype for explaining the topic similarity. The text-based interface (E2-1) simply says "You and [the scholar] have common research topics on [T1], [T2], [T3]."

E2-2: Topical Words This interface followed the approach of [90], which attempted to help users in interpreting the topic by presented topical words in a table. I adopted the idea as *E2-2 Topical Words* that present the topical words in two multi-column tables (each column contains the top 10 words of each topic). The design is shown in Figure 16a.

E2-3: FLAME This interface followed [149], which adopted a bar chart and two word

(a) E2-2: Topical Words

(b) E2-3: FLAME



(c) E2-4: Topical Radar

(d) E2-5: Topical Bar

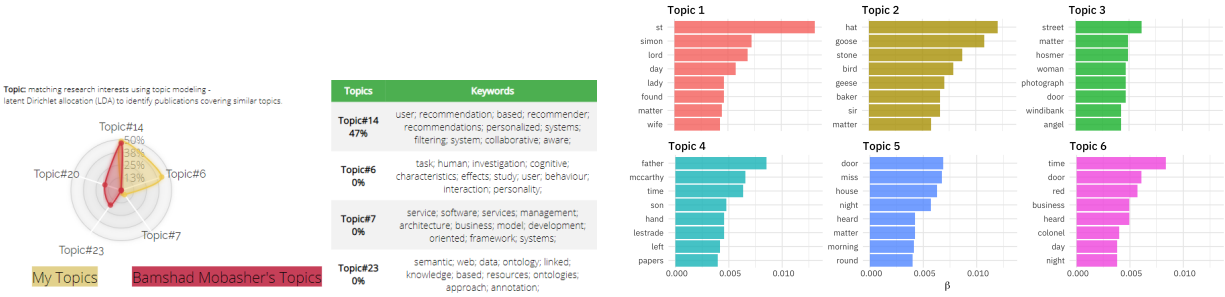


Figure 16: The prototype interfaces for *Topic Similarity* in study 2.

clouds in displaying the opinion mining result. In their design, each bar would be considered as a "sentiment"; then, the user can interpret the model by the figure (for the *beta* value of topic) and table (for the topical words). I extended the idea as *E2-3: FLAME* that showed two sets of research topics (top 5) and the relevant topic words in two word clouds (one for each scholar). The design is shown in Figure 16b.

E2-4: Topical Radar The *E2-4 Topical Radar* was used in [136]. The radar chart was presented on the left. I picked the top 5 topics (ranked by *beta* value from a total of 30 topics) of the user and compared them with the examined attendee through the overlay. A table with topical words was presented in the right so that the user can inspect the context of each research topic. The design is shown in Figure 16c.

E2-5: Topical Bars I adopted several bar charts in this interface as *E2-5: Topical Bar*. The interface showed the top three topics of two scholars (top row and the second row) and the topical information (top eight topical words in the y-axis and topic *beta* value in x-axis)

using a bar chart with histograms. The design was shown in Figure 16d.

Results The card-sorting result was presented in Table 8. I found the *E2-4 Topical Radar* received 86 votes in Rank 1 outperforming all other interfaces. E2-3 ended up being second, with most votes in the R2 group. According to the post-session interview, 13 subjects agreed E2-4 is the best interface among all examined interfaces. One subject preferred E2-3, and one subject suggested a mix of E2-3 and E2-4 as the best design. The supporting reasons for E2-4 can be summarized as 1) It is easy to see the relevance through the overlapping area from the Radar chart and the percentage numbers from the table (N=12). 2) It is informative to compare the shared research topics and topical words (N=9).

6.1.3 Explaining Co-Authorship Similarity

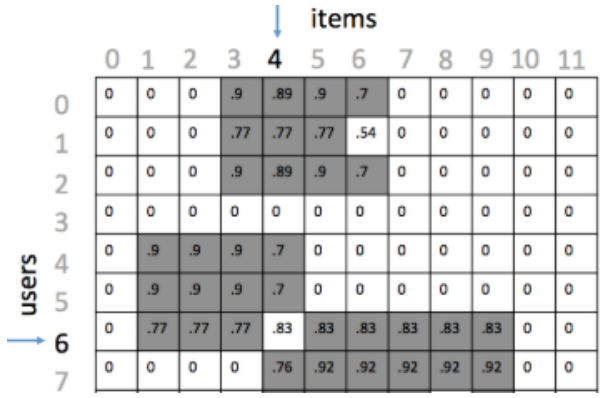
The key component of co-authorship similarity is *coauthors*, *coauthorship* and *distance of connections* of the scholars as well as its mutual relationship (i.e., the connecting path) between two scholars. I presents the five prototyping interfaces (shown in Figure 17, E3-1 presented in text below) for explaining publication similarity. In addition to four visualized interfaces, I also include one text-based interface (E3-1). That is, "You and [the scholar] have common co-authors, they are [A1], [A2], [A3]."

E3-2: Correlation Matrix *E3-2 Correlation matrix* was inspired by [57] that was used to present overlapping user-item co-clusters in a scalable and interpretable product recommendation model. I extended the interface to a user-to-user correlation matrix that the user can inspect the scholar co-authorship network. The design was shown in Figure 17(a).

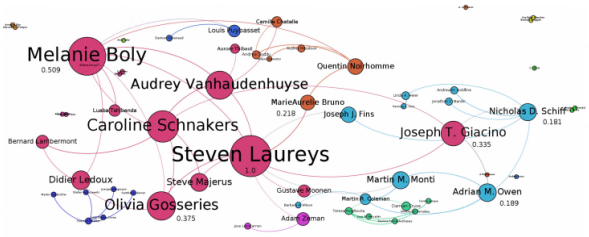
E3-3: ForceAtlas2 *E3-3: ForceAtlas2* was inspired by [38] that presented Co-authorship graph of NiMCS and related research with both high and low-level network structure and information. Nodes and edges are representing authors and co-authorship, respectively. Graph layout uses the ForceAtlas2 algorithm [38]. Clusters are calculated via Louvain modularity and delineated by color. The frequency of co-authorship is calculated via Eigenvector centrality and represented by size. The design was shown in Figure 17(b).

E3-4: Strength Graph *E3-4 Strength Graph* was inspired by [131] that tried to present

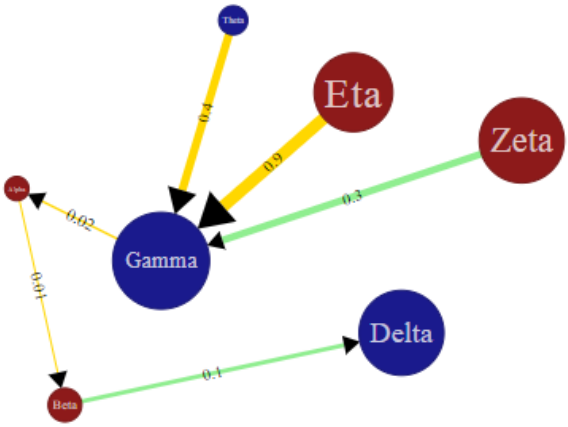
(a) E3-2: Correlation Matrix



(b) E3-3: ForceAtlas2



(c) E3-4: Strength Graph



(d) E3-5: Social Viz

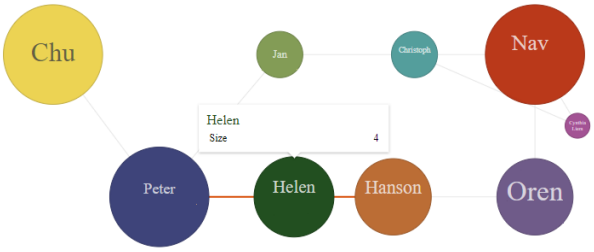


Figure 17: The prototype interfaces for *Co-Authorship Similarity* in study 2.

the co-authorship network using D3plus network style [80]. Nodes and edges are representing authors and co-authorship, respectively. The edge thickness is the weighting of the co-authorship (number of co-worked papers). The node was assigned different color by their groups, i.e., the original scholar, target scholar and via scholars. The design was shown in Figure 17(c).

E3-5: Social Viz The *E3-5 Social Viz* was used in [136]. There were six possible paths (one shortest and five alternatives). The user will be presented in the left with a yellow circle. The target user will be presented in the right with a red color. The circle size represented

the weighting of the scholar, which was determined by the appearing frequency in the six paths. For example, the scholar *Peter* is the only node that scholar *Chu* can reach scholar *Nav*, so the circle size was the largest one (size = 6). The design was shown in Figure 17(d).

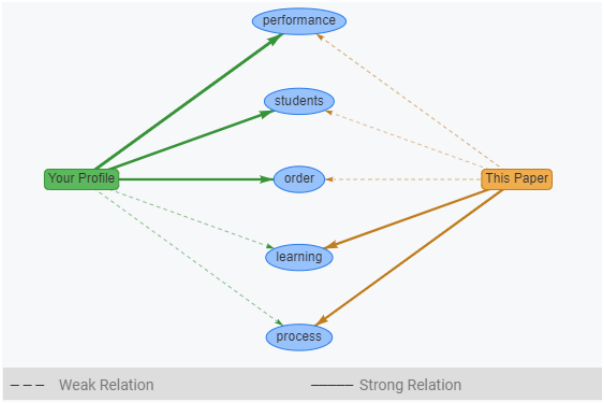
Results The card-sorting result was presented in Table 8. I found the *E3-4 Strength Graph* was preferred by the participants, received 45 votes in Rank 1. However, the votes were close with *E3-2 Correlation Matrix* (37 votes) and *E3-3 ForceAtlas2* (32 votes). According to the post-session interview, four subjects agreed E3-4 is the best interface versus the other four interfaces. The supporting reasons were the interface highlighted the mutual relations and let the user can understand the path between two scholars. The arrow and edge thickness were also useful. Two subjects supported E3-2; they liked the correlation matrix provided a clear number and correlation information that easier for them to process. Three subjects supported E3-3; they preferred the interface provided a piece of high-level information by giving a "big picture". Also, E3-3 would be good to explore the co-authorship network beyond the connecting path, although the interface was reported to be too complicated as an explanation. Four subjects supported E3-5, they enjoy the simple, clear, and "straightforward" connecting path as the explanation for co-authorship network.

6.1.4 Explaining CN3 Interest Similarity

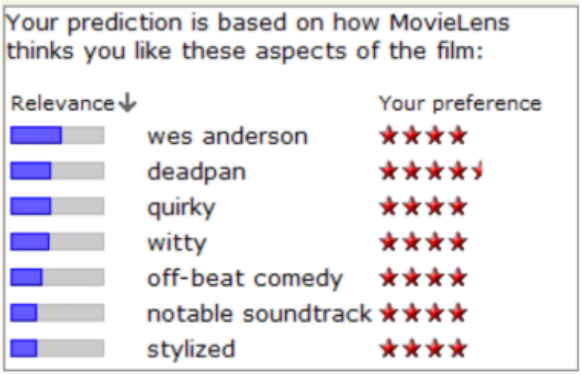
The key component of CN3 interest similarity is *papers* and *authors* of the system bookmarking as well as its mutual relationship (i.e., the common terms) between two scholars. I presented the five prototyping interfaces (shown in Figure 18, E4-1 presented in the text below) for explaining publication similarity. In addition to four visualized interfaces, I also include one text-based interface (E4-1). That is, "You and [the scholar] have common bookmarking, they are [P1], [P2], [P3]."

E4-2 Similar Keywords E4-2 Similar Keywords was proposed and deployed in Conference Navigator [109]. I extended the interface to explain shared bookmarks between two scholars. The interface represents the scholars in two sides and the common co-bookmarking items (e.g., the five common co-bookmark papers or authors) in the middle. A strong (solid line) or weak (dash line) tie will be used to connect the item was bookmarked by the one-side

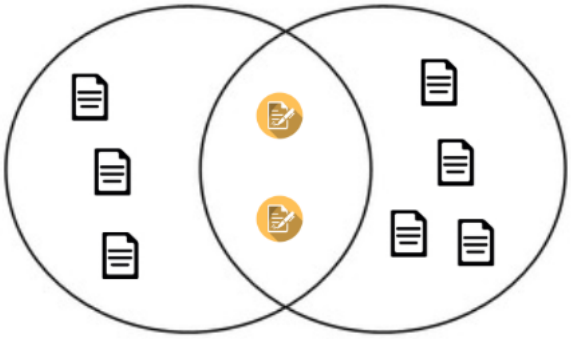
(a) E4-2: Similar Keywords



(b) E4-3: Tagsplanations



(c) E4-4: Venn Tags



(d) E4-5: Itemized List

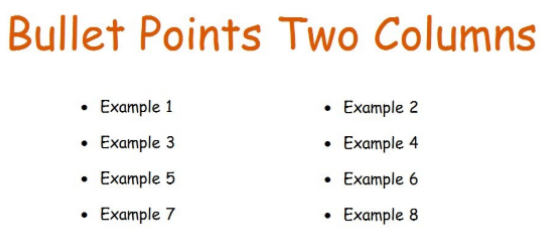


Figure 18: The prototype interfaces for *CN3 Interest Similarity* in study 2.

or two-sides. The design was shown in Figure 18(a).

E4-3: Tagsplanations *E4-3 Tagsplanations* was proposed by [141]. The idea is to show both tag, user preference, and relevance that used to recommend movies. I extended the interface to explain the co-bookmarking information. In my design, the co-bookmarked item will be listed and ranked by its social popularity, i.e., how many users have followed/bookmarked the item? The design was shown in Figure 18(b).

E4-4: Venn Tags The study of [75] has pointed out the user preferred the Venn diagram as an explanation in a recommender system. In the interface of *E4-4: Venn Tags*, I implemented the same idea with the bookmarked items. The idea is to present the book-

marked item, using an icon, in the Venn diagram. The two sides are the bookmarked item belong to one party. The co-bookmarked or co-followed item will be placed in the middle. The users can hover the icon for detail information, i.e., paper title or author name. The design was shown in Figure 18(c).

E4-5: Itemized List An itemized list has been adopted to explain the bookmark. I proposed *E4-5: Itemized List* that presented the bookmarked or followed items in two lists. The design was shown in Figure 18(d).

Results The card-sorting result was presented in Table 8. I found the *E4-4 Venn Tags* was preferred by the participants, received 64 votes in Rank 1, which was outperformed all other four interfaces. *E4-3 Tagsplanations* was also be favored by the subject, which received 49 votes. According to the post-session interview, eight subjects agreed E4-4 is the best interface versus the other four interfaces. The supporting reasons can be summarized as 1) the Venn diagram is more familiar or clear than other interfaces (N=4); 2) The Venn diagram is simple and easy to understand (N=4). Three subjects mentioned they preferred E4-3 the most due to the interface provide extra attribution, don't need to hover for detail, and easy-to-use.

6.1.5 Explaining Geographic Similarity

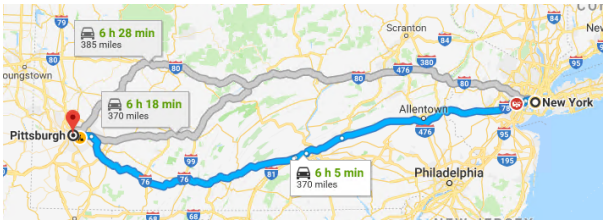
The key component of geographic similarity is *location* and *distance* of the two scholars as well as their mutual relationship (i.e., the geographic distance). I presented the five prototyping interfaces (shown in Figure 19, E5-1 presented in the text below) for explaining the geographic similarity. In addition to four visualized interfaces, I also include one text-based interface (E5-1). That is, "From [Institution A] to [sample]'s affiliation ([Institution B]) = N miles."

E5-2: Earth Style Using Google Map [61] for explaining geographic distance in a social recommender system has been discussed in [**?**]. I extended the interface to a different style. In *E5-2 Earth Style*, I "zoom out" the map to an earth surface and place the two connected icons (with geographic distance) on the map. The design was shown in Figure 19(a).

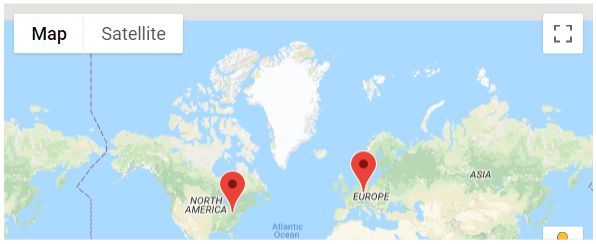
E5-3: Navigation Style *E5-3 Navigation Style* followed the same Google Map API

(a) E5-2: Earth Style

(b) E5-3: Navigation Style



(c) E5-4: Icon Style



(d) E5-5: Label Style

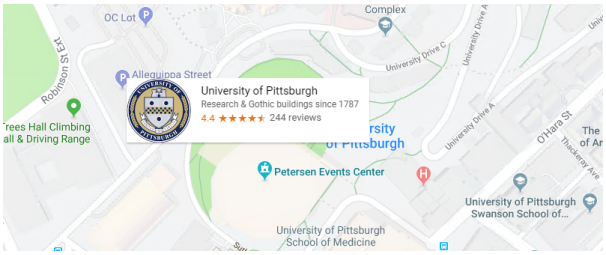


Figure 19: The prototype interfaces for *Geography Similarity* in study 2.

(shown in E5-2), but presented navigation between the two locations, either by car or flight. To be noted, the *transportation time*, i.e., the fly or driving time in E5-2 or E5-3, did not be considered in the recommendation model. The design was shown in Figure 19(b).

E5-4: Icon Style *E5-4 Icon Style* followed the same Google Map API (shown in E5-2), but presented two icons on the map without any navigation information. The users can hover to see the detail affiliation, but the geographic distance information was not presented. The design was shown in Figure 19(c).

E5-5: Label Style *E5-4 Label Style* followed the same Google Map API (shown in E5-2), but presented two labels on the map without any navigation information. The users can see the detail affiliation profile through the floating label without extra clicking or hovering interactions. The design was shown in Figure 19(d).

Results The card-sorting result was presented in Table 8. I found the *E5-3 Navigation Style* was preferred by the participants, received 42 votes in Rank 1. However, the votes are close with *E5-3 Label Style* (40 votes). In the post-session interview, six subjects agreed E5-3

is the best interface versus the other four interfaces. Three subjects particularly mentioned the navigation function was irrelevant in explaining or exploring the social recommendations. The supporting reasons for E5-3 can be summarized as 1) The map is informative (N=2). 2) It is useful to see navigation (N=5). Three subjects mentioned they preferred E5-5 the most due to the label contains affiliation information that they can understand the affiliation without extra actions. Although there is no geographic distance information, one subject pointed out he will realize the distance after knowing the affiliation title.

6.1.6 Summary and Discussion

In this stage, I proposed a total of 25 explanation interfaces for five recommendation models and reported the card-sorting and semi-interview results. I found, in general, the participants preferred visualization interfaces more than the text-based interface. Based on study 2, I found *E1-4: Venn Word Cloud*, *E2-4: Topical Radar*, *E3-4: Strength Graph*, *E4-4: Venn Tags*, *E5-3: Navigation Style* were preferred by the study participants. I further discussed the top-rated and second-rated explanation interfaces and user feedback in each session. Based on the experiment results, I concluded the design implication of bringing the explanation interface to a real-world social recommender system.

I choose the explanation interface design using seven factors. The seven factors included Transparency (TP), Scrutability (SC), Trust (TS), Persuasiveness (PE), Effectiveness (ET), Efficiency (EF), and Satisfaction (SA). The detailed statement of each factor can be found in Table 9. In this dissertation, I select the user preferred explanation interface design by the votes in the *first rank*, i.e., the preferred interface design in general. However, it is also possible to elaborate on the design selection by a deeper analysis of the card-sorting results. That is, a recommender system may aim to enhance one particular exploratory goal. For example, instead of selecting the general preferred design, the researcher or designer may select an interface design that can *persuade* users to accept the recommendation. In this case, factor 12, 13, 14 should be assigned with higher weightings. Another example is to allow the users to tell if the recommendation is correct, i.e., to enhance the scrutability of the system. In this case, the factor of 6, 7, 8, and 9 would better reflect the design implications.

Table 8: The card-sorting results of study 2.

	R1	R2	R3	R4	R5	Not Applicable	Total Votes
E1-1	19	25	21	19	44	22	150
E1-2	23	37	17	30	26	17	150
E1-3	7	16	42	44	19	22	150
E1-4	**76**	32	27	2	0	13	150
E1-5	19	31	33	28	20	19	150
E2-1	12	8	14	21	60	35	150
E2-2	6	2	9	73	36	24	150
E2-3	24	78	28	7	2	11	150
E2-4	**86**	31	13	11	0	9	150
E2-5	13	21	70	14	11	21	150
E3-1	13	5	9	18	69	36	150
E3-2	37	26	17	36	20	14	150
E3-3	32	38	29	28	11	12	150
E3-4	**45**	41	37	11	0	16	150
E3-5	15	32	41	36	11	15	150
E4-1	8	11	6	31	64	30	150
E4-2	17	61	48	16	2	6	150
E4-3	49	41	41	11	3	5	150
E4-4	**64**	28	41	7	1	9	150
E4-5	8	5	6	65	46	20	150
E5-1	20	7	13	24	55	31	150
E5-2	16	22	6	45	36	25	150
E5-3	**42**	16	44	11	6	31	150
E5-4	15	49	36	18	4	28	150
E5-5	40	35	26	20	3	26	150

6.2 SECOND STAGE: FIRST-ROUND EVALUATION (STUDY 3)

This section presents study 3, which I conducted to develop and evaluate explanation interfaces for five similarity-based people recommendation models. In study 2 (N=15), I introduced a card-sorting task to identify the user preferred explanation interface design for multiple explanation goals. I assessed a total of twenty-five explanation interfaces and selected the top-voted interface designs for each similarity-based recommendation model. In study 3 (N=18), I used a performance-focused evaluation approach to investigate whether using two types of explanations in parallel could offer an advantage over a single type of explanation. I compared ten explanation interfaces, one baseline plus one enhanced version for each of the five recommendation models. In each case, I use the top-rated interface as a baseline for each model and a combination of first and second preferred interfaces as an enhanced version. I implemented ten explanation interfaces using the data from a state-of-the-art scholarly social recommender system. I evaluated the explanation interfaces by asking the participant to "sort" recommendation based on the relevance. I analyzed the findings combined with the sorting result, user perception survey, and NASA-TLX survey.

6.2.1 Similarity-Based Recommendations

In this section, I present the results of my attempts to design and evaluate visual explanations for five similarity-based people recommendation models: *text similarity*, *topic similarity*, *co-authorship similarity*, *CN3 interest similarity*, and *geography similarity*. These models are widely adopted in many content-based recommender systems [132, 136, 141].

Text similarity (E1) is a metric that measures similarity or dissimilarity (distance) between two text strings [43]. The "strings" can consist of various information sources. For example, in a scholarly people recommender system, the string can be generated from scholars' academic publications. To measure the text-similarity (distance), one promising approach is to convert the strings into a *term vectors* and then compute their *cosine similarity* [132]. A higher similarity (i.e., the shortest distance) between "strings" representing publications of two researchers indicates that the two researchers have a larger fraction of

common terms in the text of their publications.

Topic similarity (E2) is a metric that measures the distance between topic distributions [31]. This is another approach to measure the similarity between the publications of two researchers. The approach assumes that a mixture of topics is used to generate a string (document), where each topic is a distribution of topical words. A social recommender engine, based on the topic-based approach, can represent the scholars' research interests through the learned *topics*. The topic similarity could be computed as the pairwise similarity of the topic distributions [136]. In my study, the *topics* were generated by topic modeling, Latent Dirichlet Allocation (LDA), by classifying their publication text [31]. A higher topic similarity means a shorter distance between the two scholars' research interests, i.e., the two scholars shared more common research topics.

Co-authorship similarity (E3) is a metric that measures the connectivity of two vertices in a social network. The connectivity can be defined by different measurements, for example, the number of common neighborhood, number of paths or shortest distance [121, 95, 132]. In a scholar recommender system, the co-authorship similarly can be calculated by measuring the *distance* between two scholars, based on their co-authorship network. A higher social similarity means that the two scholars have a shorter distance in their co-authorship network, i.e., the two scholars are connected with a fewer node degree.

CN3 interest similarity (E4) is a metric that measures the portion of shared items, which can be varied in a different context. For example, items shared by two users could be user-generated tags [141] or friends followed on social media. In a scholar recommender system, the item similarity between two scholars can be calculated by measuring the intersection of the bookmarked papers [10, 81], e.g., using Jaccard similarity [132, 136]. A higher item similarity means that the two scholars have more similar interests with respect to the academic articles or conference presentation, i.e., they co-bookmarked a larger number of papers at the same conference.

Geography similarity (E5) is a metric that measures the geographic distance between two entities. For example, the distance can be determined through longitude and latitude data based on locations. In a scholar recommender system, the geography similarly can be calculated by two scholars' affiliation information, The affiliation location can be converted

in a coordinate system, and then applied the Haversine formula to compute the geographic distance between scholars [138]. A higher geography similarity means that the two scholars are affiliated with the nearby institutions, i.e., the two scholars may live or work in the nearby cities or regions.

6.2.2 Developing Explanation Interfaces

I designed visual interfaces to explain five similarity-based models for recommending conference attendees to meet implemented in a conference support system Conference Navigator (CN) [10]. All interfaces were selected to visually explain one type of "similarity" between the user and a recommended scholar described in the previous section. My goal at this stage was to find visualizations that can better explain the similarity model as measured by user perception. Existing state-of-the-art explanation interfaces or models motivated my interface designs. Due to the page limit, this paper shows only top-performing designs (see Figure 20). The full set of designs can be found in [134].

6.2.2.1 Explaining Text Similarity (E1) The key component of text similarity is *terms* and *term frequency* of the publication as well as its mutual relationship (i.e., the common terms) between two scholars. I presented one text-based interface (**E1-1**) and four visual interfaces (**E1-2** to **E1-5**) for explaining text similarity.

E1-1 Text-Based Explanation: The text-based interface was presenting the explanation as: *You and [the scholar] have common words in [W1], [W2], [W3].*

(Second-rated) E1-2 Two-way Bar Chart: The bar chart is a common approach in analyzing the text mining outcome using a histogram of terms and term frequency [118]. I extended the design to a two-way bar chart to better compare two scholars' publication terms and term frequency, i.e., one scholar on the right and the other scholar on the left (Figure 20b).

E1-3 Word Clouds: Word cloud is a universal design in explaining text similarity [40, 131]. I adopted the word cloud style from [149], which presented the term in the cloud and the term frequency by the font size. I used two-word clouds (one for each scholar), so

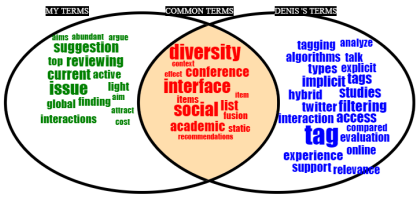
Table 9: The explanation factors of study 3

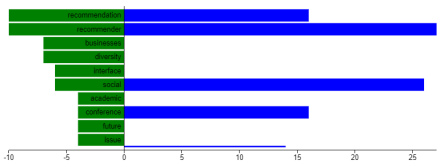
	Factor	Statement
1	TP, SA	The visualization presents the similarity between my interest and the recommended person.
2	TP	The visualization presents the relationship between the recommended person and me.
3	TP	The visualization presents where the data was retrieved.
4	TP	The visualization presents more in-depth information on how the scores sum up.
5	TP, TS ET	The visualization allows me to see the connections between people and understand how they are connected.
6	SC	The visualization allows me to understand whether the recommendation is good or not.
7	SC	The visualization presents the data for making the recommendations.
8	SC	The visualization allows me to compare and decide whether the system is correct or wrong.
9	SC	The visualization allows me to explore and then determine the recommendation quality.
10	TS	The visualization presents a convincing explanation to justify the recommendation.
11	TS	The visualization presents the components (e.g., algorithm) that influenced the recommendation.
12	PE	The visualization shows me the shared interests, i.e., why my interests are aligned with the recommended person.
13	PE, SA	The visualization has a friendly, easy-to-use interface.
14	PE	The visualization inspired my curiosity to discover more information.
15	ET	The visualization presents the recommendation process clearly.
16	EF	The visualization presents highlighted items or information that is strongly related to me.
17	EF	The visualization presents aggregated, non-obvious relations to me.
18	SA	The visualization presents feedback from other users, i.e., I can see how others rated a recommended person.
19	SA	The visualization allows me to tell why does this system recommend the person to me.

the user can perceive the mutual relationship.

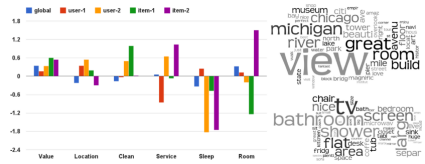
(Top-rated) E1-4 Venn Word Cloud: This interface could be considered as a combination of a word cloud and a Venn diagram [136], which presents term frequency using the

(a) E1-4: Venn Word Cloud (Top-rated) **(b) E1-2**: Two-way Bar Chart (Second-rated)

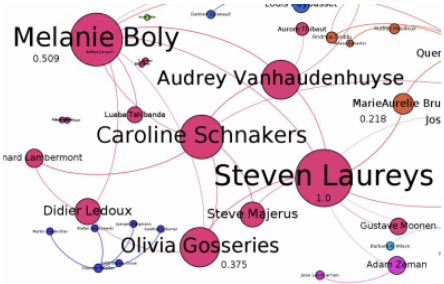




(c) E2-4: Topical Radar (Top-rated) **(d) E2-3**: FLAME (Second-rated)

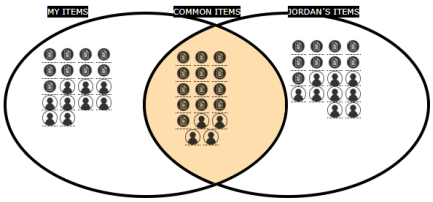


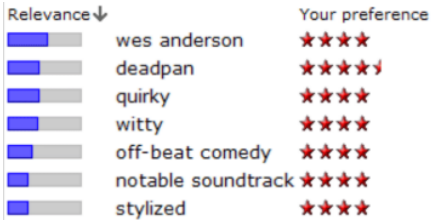
(e) E3-3: ForceAtlas2 (Top-rated) **(f) E3-2**: Correlation Matrix (Second-rated)



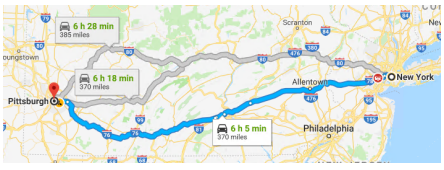
0	0	0	.9	.89	.9	.7	0	0	0	0
0	0	0	.77	.77	.77	.54	0	0	0	0
0	0	0	.9	.89	.9	.7	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0
0	.9	.9	.9	.7	0	0	0	0	0	0
0	.9	.9	.9	.7	0	0	0	0	0	0
0	.77	.77	.77	.83	.83	.83	.83	.83	.83	0
0	0	0	0	.76	.92	.92	.92	.92	.92	0

(i) E4-4: Venn Tags (Top-rated) **(j) E4-3**: Tagsplanations (Second-rated)





(k) E5-3: Navigation Style (Top-rated) **(l) E5-5**: Label Style (Second-rated)



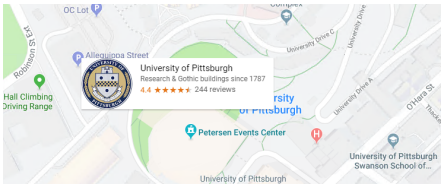


Figure 20: The top-rated and second-rated visual interfaces study 3.

font size. The unique terms of each scholar are shown in a different color (green and blue) while the common terms are presented in the middle, with red color, for determining the mutual relationship (Figure 20a).

E1-5 Interactive Word Cloud: A word cloud can be interactive. I extend the idea from [131] and used "Zoomdata Wordcloud" tool [154], which follows the common approach to visualize term frequency with the font size. The term color was selected to distinguish the scholars' terms, i.e., different term colors for each scholar. A slider was attached to the bottom of the interface that provides real-time interactive functionality to increase or decrease the number of terms in the word cloud.

6.2.2.2 Explaining Topic Similarity (E2)

The key component of topic similarity is *research topics* and *topical words* of the scholar as well as its mutual relationship (i.e., the common research topics) between two scholars. I presented one text-based interface (**E2-1**) and four visual interfaces (**E2-2** to **E2-5**) for explaining topic similarity.

E2-1 Text-Based Explanation: The interface was presenting the explanation as *You and [the scholar] have common research topics on [T1], [T2], [T3].*

E2-2 Topical Words: This interface extended the approach by [90], which attempted to enable topic interpretation by presenting topical words in a table. I adopted the idea as *E2-2 Topical Words* that present the topical words in two multi-column tables (each column contains ten topical words).

(Second-rated) E2-3 FLAME: This interface was proposed by [149], which adopted a bar chart and two word-clouds in displaying the topical mining result. The user can interpret the topic model by the diagram (for the *beta* value of the topic) and the table (for the topical words). I extended the idea as *E2-3: FLAME* that showed two sets of research topics (top 5) and the relevant topic words in two word-clouds (one for each scholar). (Figure 20d)

(Top-rated) E2-4 Topical Radar: The interface was introduced by [136], which presented a radar diagram with a topical word table. The radar filed the top 5 topics (ranked by *beta* value) of the user and the corresponding value of the recommended scholar. The table with topical words was presented in the right so that the user can inspect the context of each research topic. (Figure 20c)

E2-5 Topical Bars: Bar chart have been shown useful in analyzing the frequency across different topics [118]. In this interface, I adopted multiple bar charts to show six topic distribution of two scholars and the topical information (words and topic *beta* value).

6.2.2.3 Explaining Co-Authorship Similarity (E3)

The key component of co-authorship similarity is *coauthors*, *coauthorship* and *distance of connections* of the scholars as well as its mutual relationship (i.e., the connecting path) between two scholars. I adopt one text-based interface (**E3-1**) and four visual interfaces (**E3-2** to **E3-5**) for explaining topic similarity.

E3-1: Text-Based Explanation: The interface was presenting the explanation as "You and [the scholar] have common co-authors, they are [A1], [A2], [A3]."

E3-2: Correlation Matrix *E3-2 Correlation matrix* was inspired by [57] that was used to present overlapping user-item co-clusters in a scalable and interpretable product recommendation model. I extended the interface to a user-to-user correlation matrix that the user can inspect the scholar co-authorship network (Figure 20f).

(Second-rated) E3-3: ForceAtlas2 *E3-3: ForceAtlas2* was inspired by [38] that presented Co-authorship graph of NiMCS and related research with both high and low-level network structure and information. Nodes and edges are representing authors and co-authorship, respectively. Graph layout uses the ForceAtlas2 algorithm [38]. Clusters are calculated via Louvain modularity and delineated by color. The frequency of co-authorship is calculated via Eigenvector centrality and represented by size (Figure 20e).

(Top-rated) E3-4: Strength Graph *E3-4 Strength Graph* was inspired by [131] that tried to present the co-authorship network using D3plus network-style [80]. Nodes and edges are representing authors and co-authorship, respectively. The edge thickness is the weighting of the co-authorship (number of co-worked papers). The node was assigned different colors by their groups, i.e., the original scholar, target scholar, and scholars.

E3-5: Social Viz The *E3-5 Social Viz* was used in [136]. There were six possible paths (one shortest and five alternatives). The user will be presented in the left with a yellow circle. The target user will be presented in the right with a red color. The circle size represented the weighting of the scholar, which was determined by the appearing frequency in the six paths. For example, the scholar *Peter* is the only node that scholar *Chu* can reach scholar

Nav, so the circle size was the largest one (size = 6).

6.2.2.4 Explaining Item Similarity (E4) The key component of item similarity is the *papers* and *authors* of the bookmarking as well as its mutual relationship (i.e., the common items) between two scholars. I present the five prototyping interfaces for explaining item similarity. In addition to four visualized interfaces (E3-2 to E3-5), I also include one text-based interface (E3-1) for explaining item similarity.

E4-1 Text-Based Explanation: The interface was presenting the explanation as: *You and [the scholar] have common bookmarking, they are [B1], [B2], [B3].*

E4-2 Similar Keywords: The interface was proposed and deployed by the CN system [109]. I extended the interface to explain common bookmarks between two scholars. The interface represents the scholars in two sides and the common co-bookmarking items (e.g., the five common co-bookmark papers or authors) in the middle. A strong (solid line) or weak (dash line) tie will be used to connect the item was bookmarked by the one-side or two-sides.

(Second-rated) E4-3: Tagsplanations: The visualization was proposed by [141]. The idea is to show tag and relevance in an ordered bar. I extended the interface to explain the co-bookmarking information. In my design, the co-bookmarked item will be listed and ranked by its social popularity, i.e., how many users have followed/bookmarked the item? (Figure 20j)

(Top-rated) E4-4: Venn Tags: The study [75, 104] pointed out that users preferred Venn diagrams as a way to explain recommendation. In the interface, presented items bookmarked by compared scholars as icons on the Venn diagram. Two sides of the diagram show bookmarked item belonging only to one of the compared scholars. The co-bookmarked items are presented in the middle. The users can mouse over the icon for detail information, i.e., paper title. (Figure 20i)

E4-5: Itemized List: An itemized list has been adopted to explain the bookmark in [**?**]. I extended the design in presenting the bookmarked or followed items in two comparable itemized lists.

6.2.2.5 Explaining Geographic Similarity (E5) The key component of geographic similarity is *location* and *distance* of the two scholars as well as their mutual relationship (i.e., the geographic distance). I presented the five prototyping interfaces (shown in Figure **??**, E5-1 presented in the text below) for explaining the geographic similarity. In addition to four visualized interfaces, I also include one text-based interface (E5-1). That is, "From [Institution A] to [sample]'s affiliation ([Institution B]) = N miles."

E5-2: Earth Style I adopt Google Map [61] for explaining geographic distance in a social recommender system. I extended the interface to a different style. In *E5-2 Earth Style*, I "zoom out" the map to an earth surface and place the two connected icons (with geographic distance) on the map.

(Top-rated) E5-3: Navigation Style *E5-3 Navigation Style* followed the same Google Map API (shown in E5-2), but presented navigation between the two locations, either by car or flight. To be noted, the *transportation time*, i.e., the fly or driving time in E5-2 or E5-3, did not be considered in the recommendation model (Figure 20k).

E5-4: Icon Style *E5-4 Icon Style* followed the same Google Map API (shown in E5-2), but presented two icons on the map without any navigation information. The users can hover to see the detail affiliation, but the geographic distance information was not presented.

(Second-rated) E5-5: Label Style *E5-4 Label Style* followed the same Google Map API (shown in E5-2), but presented two labels on the map without any navigation information. The users can see the detail affiliation profile through the floating label without extra clicking or hovering interactions (Figure 20l).

6.2.2.6 Card-Sorting Analysis The card-sorting results are presented in Table 10. The first (top-rated) and the second (second-rated) most-preferred interfaces are highlighted in red and blue color, respectively. In general, the result indicated that the participants preferred visual explanations over text-based explanations. The pattern was consistent in all five tasks; the text-based explanations were always received the most *Group 5* and *Not Applicable* votes.

In explaining text similarity (E1), *E1-4 Venn Word Cloud* was preferred by the participants (received 117 votes in Group 1) outperforming the other four interfaces. According

Table 10: The card-sorting results of study 3

	R1	R2	R3	R4	R5	Not Applicable	Total Votes
E1-1	32	39	30	30	76	78	285
E1-2	**47**	55	31	56	35	61	285
E1-3	12	33	68	68	34	70	285
E1-4	**117**	57	44	8	2	57	285
E1-5	38	50	56	45	31	65	285
E2-1	18	9	23	29	113	93	285
E2-2	18	18	22	119	49	59	285
E2-3	**48**	125	56	13	2	41	285
E2-4	**137**	55	26	25	2	40	285
E2-5	30	39	108	35	13	60	285
E3-1	17	6	15	24	127	96	285
E3-2	**74**	46	38	58	25	4	285
E3-3	**74**	71	42	44	14	40	285
E3-4	67	76	71	21	1	49	285
E3-5	20	51	70	73	17	54	285
E4-1	11	13	6	48	113	94	285
E4-2	32	103	87	22	2	39	285
E4-3	**91**	75	66	14	4	35	285
E4-4	**101**	48	68	15	2	51	285
E4-5	17	10	12	116	63	67	285
E5-1	45	12	16	41	93	78	285
E5-2	25	36	13	77	59	75	285
E5-3	**75**	37	65	16	9	83	285
E5-4	23	74	65	39	6	78	285
E5-5	**59**	63	55	28	6	74	285

to the post-stage interview, 13 subjects agreed that E1-4 is the best interface due to the following reasons. 1) the Venn diagram provided common terms in the middle, which highlighted the common terms and shared relationship; 2) it is useful to show non-overlapping terms on the sides (N=5) and 3) the design is simple, easy to understand and require less time to process (N=3). Two subjects particularly mentioned they preferred *E1-2: Two-way Bar Chart* (received 47 votes in Group 1, second-rated interface) since histograms give them

the "concrete numbers" for "calculating" the similarity, which was harder when using word clouds.

In explaining topic similarity (E2), *E2-4 Topical Radar* received 137 votes in Group 1, outperforming all other interfaces. *E2-3 FLAME* ended up second in Group 1 as well as received the most votes in Group 2. According to the post-stage interview, 13 subjects agreed E2-4 is the best interface among all examined interfaces. The supporting reasons for E2-4 can be summarized as 1) It is easy to see the relevance through the overlapping area from the radar chart and the percentage numbers from the table (N=12). 2) It is informative to compare the shared research topics and topical words (N=9). One subject specifically preferred E2-3, and one subject suggested a mix of E2-3 and E2-4 as the best design.

In explaining co-authorship similarity (E3), *E3-3: ForceAtlas2* and *E3-2: Correlation Matrix* both received 74 votes in Group 1. The card-sorting result is surprising due to the two interfaces were explaining co-authorship similarity in a different way. When the votes in Group 2 were considered, the social network-style interfaces (E3-3: ForceAtlas2 and E3-4: Strength Graph) were still favored more by the users. According to the post-stage interview, user preference was diverse. There were three subjects preferred E3-3 due to the simplicity of recognizing the big picture as well as the detailed information. There were three subjects supported E3-4 due to the better usability of only showing the relevant information. Two subjects top-ranked E3-2 due to their familiarity and the presenting of correlation. One subject particularly mentioned she preferred E3-1 due to all visualize interfaces were too complicated for her.

In explaining item similarity (E4), *E4-4 Venn Tags* received 101 votes in Group 1 outperforming the other four interfaces. *E4-3 Tagsplanations* finished as a very close second receiving 91 votes. According to the post-stage interview, eight subjects agreed that E4-4 is the best interface among the five interfaces. The supporting reasons can be summarized as 1) the Venn diagram is more familiar or clear than other interfaces (N=4); 2) The Venn diagram is simple and easy to understand (N=4). Three subjects particularly mentioned that they preferred E4-3 since this interface provides extra information without requiring extra Mouse-hovering efforts while inspecting the details.

In explaining geography similarity (E3), *E5-3: Navigation Style* received 75 votes in

Group 1. *E5-5: Label Style* ended up second in Group 1 with 59 votes. According to the post-stage interview, six subjects supported E5-3 is the best design. The supporting reasons included it was clear to show the distance, navigation, and routes. To be noted, the navigation information was irrelevant to the recommendation engine, but the participants were still in favor of the interface design. Three subjects supported the design of E5-5; they agreed the interface is showing clear, detailed affiliation information. Although the distance information was missing, one subject particularly mentioned he could "appraise" the distance based on the affiliation. Three subjects preferred E5-2 due to the viability of the global map and locations. One subject chosen E5-1 due to it is less complexity of getting the distance information.

6.2.3 Assessing Visual Enhanced Explanations

Based on the card-sorting result of study 1b, I implemented the top-rated designs to assess visual explanation interfaces more reliably. The screenshot of the most-preferred interfaces can be found in Figure 20. I proposed 1) *E1-4 Venn Word Cloud* to explain text similarity (**Sim1**), 2) *E2-4 Topical Radar* to explain topic similarity (**Sim2**), 3) *E3-4: Strength Graph* to explain co-authorship similarity (**Sim3**), 4) *E4-4: Venn Tags* to explain item similarity (**Sim4**) and 5) *E5-3: Navigation Style* to explain geography similarity (**Sim5**). for the visualizations.

At the same time, the result of the post-stage interview indicated that while second-rated interfaces collected fewer votes than top-rated interfaces, participants mentioned different reasons for preferring these interfaces. I hypothesized that the features of first and second design choices could complement each other and decided to explore whether I could improve the value of the most-preferred interface by enhancing it with the second-most-preferred design. That is, I added the *E1-2 Two-way Bar Chart* to *E1-4 Venn Word Cloud* to provide additional term comparison information (**Sim1+**, Figure 21), attached two word clouds to *E2-4 Topical Radar* to mix up the user preferred component of E2-3 (**Sim2+**, Figure 22), revised *E3-4 Strength Graph* of different edge thickness (**Sim3+**, Figure 23), provided an extra list to *E4-4: Venn Tags* (**Sim4+**, Figure 24) to decrease the need for mousing-over

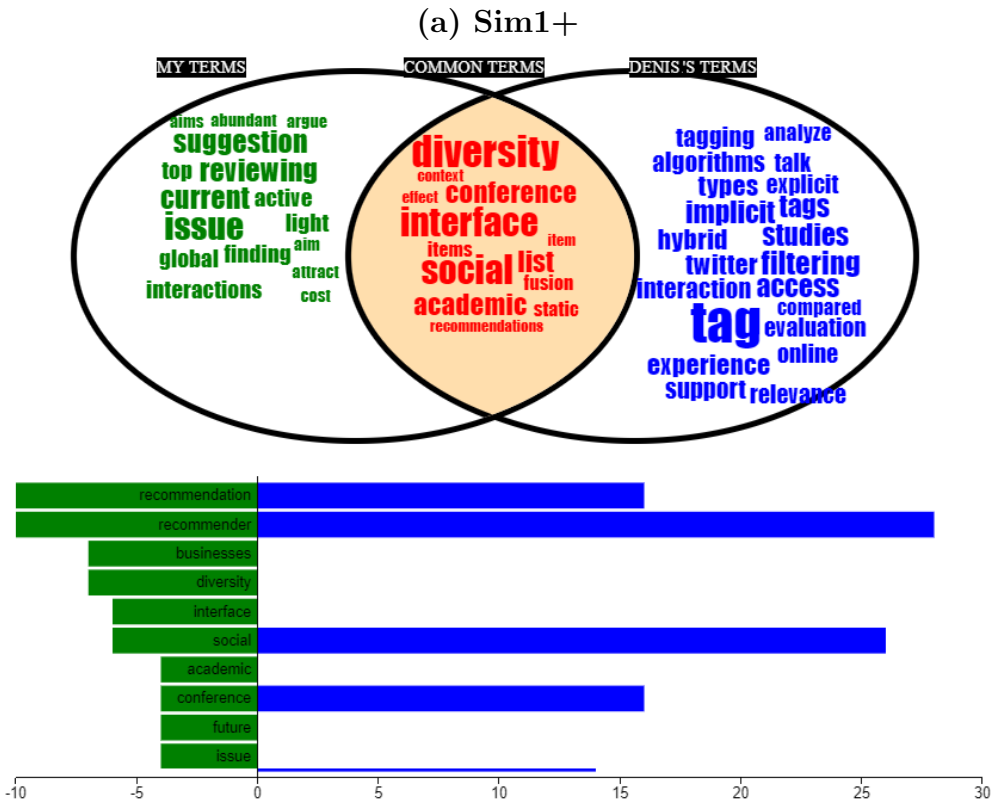


Figure 21: The visual interfaces that used to explain *publication* similarity-based recommendation model in study 3.

while getting the item details, and added location label to the design of *E5-3 Navigation Style* (**Sim5+**, Figure 25). I aimed to answer the following research questions (RQs):

- How does the visual interface reach the explanation goals?
- How does user perception vary with the enhanced interface?
- How does the explanation interface affect user performance (inspectability) across recommendations?

6.2.4 Evaluating Enhanced Explanation Interfaces

I conducted a controlled user study to evaluate and compare the selected interfaces for explaining the five similarity-based recommendation models. I introduced a total of ten ex-

(a) Sim2+

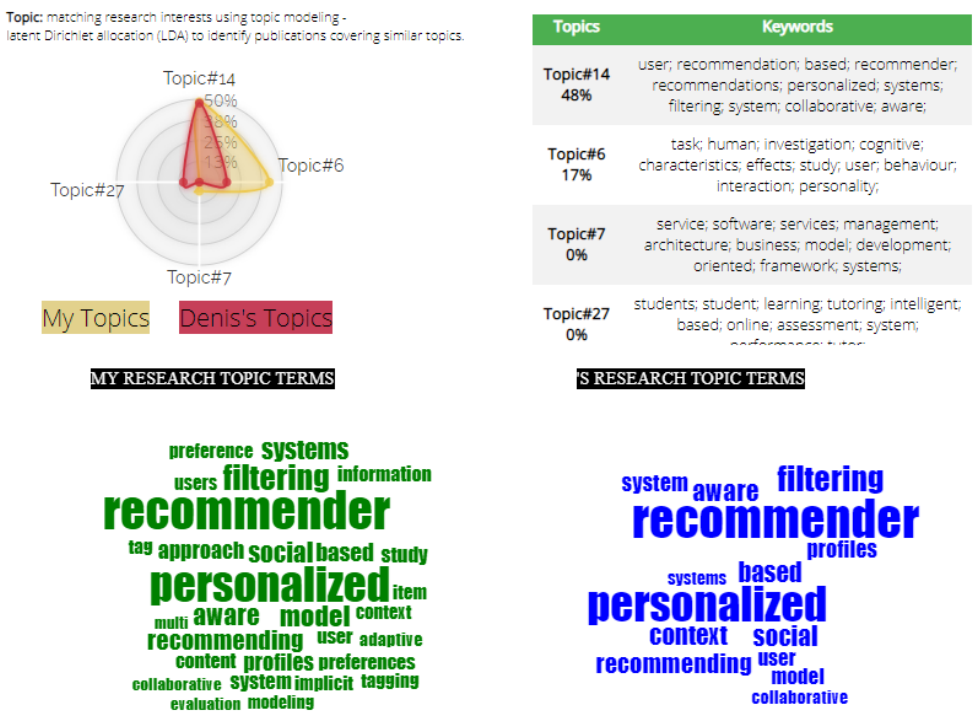


Figure 22: The visual interfaces that used to explain *topic* similarity-based recommendation model in study 3.

planation interfaces (five baseline and five enhanced interfaces) in the context of the attendee recommender component of the Conference Navigator (CN) [10]. A total of 18 (11 female) participants (N=18) were recruited for this study. There were 16 information science graduate students and two graduate from nursing and linguistics programs at the University of Pittsburgh. Their age ranged from 21 to 35 years ($M = 24.94$, $SE = 3.24$). All participants had no previous experience of using the CN system. Each participant received USD$20 compensation and signed informed consent.

In the beginning of study 3, I first introduced the CN system and the recommendation models to the subjects. After the introduction, I asked the subjects to complete a "recommendation-sorting task" using the given explanation interface, i.e., the subjects were required to *rank the recommendation relevance solely based on the visual explanation*. The

(a) Sim3+

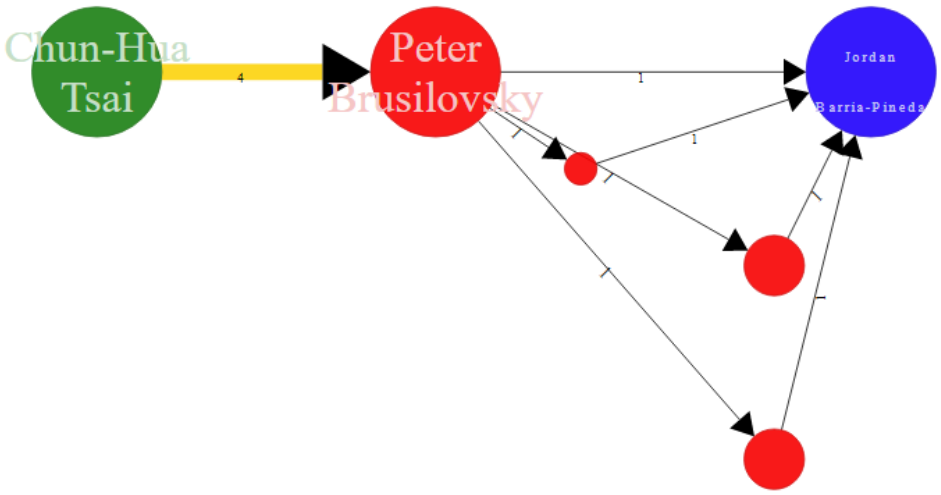
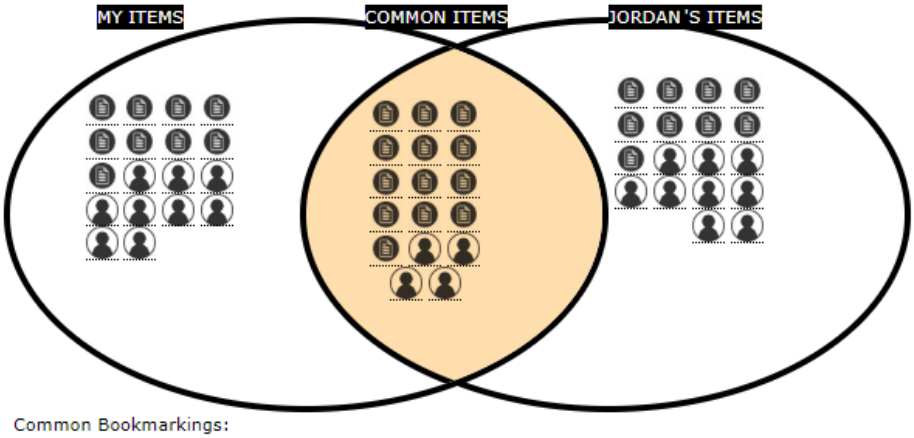


Figure 23: The visual interfaces that used to explain *co-authorship* similarity-based recommendation model in study 3.

tasks were designed to evaluate *how well an explanation interface supports the user performance of comparing the relevance across recommendations*. The experiment adopted a within-subject design, i.e., all participants were asked to perform six sorting tasks using the proposed explanation interfaces. In each task, the subject received five people recommendations generated by one recommendation model. To make the conditions equal, all users received the same recommendation generated using data of a scholar who used the CN system for at five conferences. The subjects can click the recommendation link to open the corresponding explanation interface. The five people recommendations were displayed as five links with the names of the recommended scholars. All related background information (e.g., list of publications, affiliation, title, etc.) was hidden to reduce the bias. The order of recommendation and explanation interfaces were randomized to avoid the ordering effect. To reduce the learning bias, I used data from different conferences to generate recommendations, i.e., IUI 2017, for the baseline interfaces and UMAP 2017 for the enhanced interfaces.

After each task, the subjects were asked to fill in a three-part post-stage questionnaire. First, the subjects were asked to rank the five recommendations by relevance (from high

(a) Sim4+



Common Bookmarkings:

- An Interactive and Interpretable Interface for Diversity in Recommender Systems
- Implicit Acquisition of User Personality for Augmenting Recommender Systems
- VizRec: Recommending Personalized Visualizations
- Visual Exploration and Analysis of Recommender Histories
- A 3D Item Space Visualization for Presenting and Manipulating User Preferences in Collaborative Filtering
- Cartograph: Unlocking Spatial Visualization Through Semantic Enhancement
- Improving Search Result Comprehension by Topic-Relevance Map Visualization
- CQAVis: Visual Text Analytics for Community Question Answering

Common Friends:

- Peter Brusilovsky
- Maria Bielikova
- Christoph Trattner
- Hung Chau

Figure 24: The visual interfaces that used to explain *CN3 interest* similarity-based recommendation model in study 3.

to low relevance). I measured the correct rate by *Levenshtein Distance*, i.e., given correct order as "ABCDE" and submitted answer as "ABDCE", the Levenshtein distance is 2 and the correct rate is 60% ($(5-2)/5 = 0.6$). Second, the subjects answered the nineteen-factor questions (shown in Table 9). Third, the subjects answered four NASA-TLX questions [46]. The NASA-TLX question included: *(TLX1) Mental Demand: How mentally demanding was the task? (TLX4) Performance: How successful were you in accomplishing what you were asked to do? (TLX5) Effort: How hard did you have to work to accomplish your level of performance?* and *(TLX6) Frustration: How insecure, discouraged, irritated, stressed, and annoyed were you?* The order of question was the same to all participants with a 5-point

(a) Sim5+

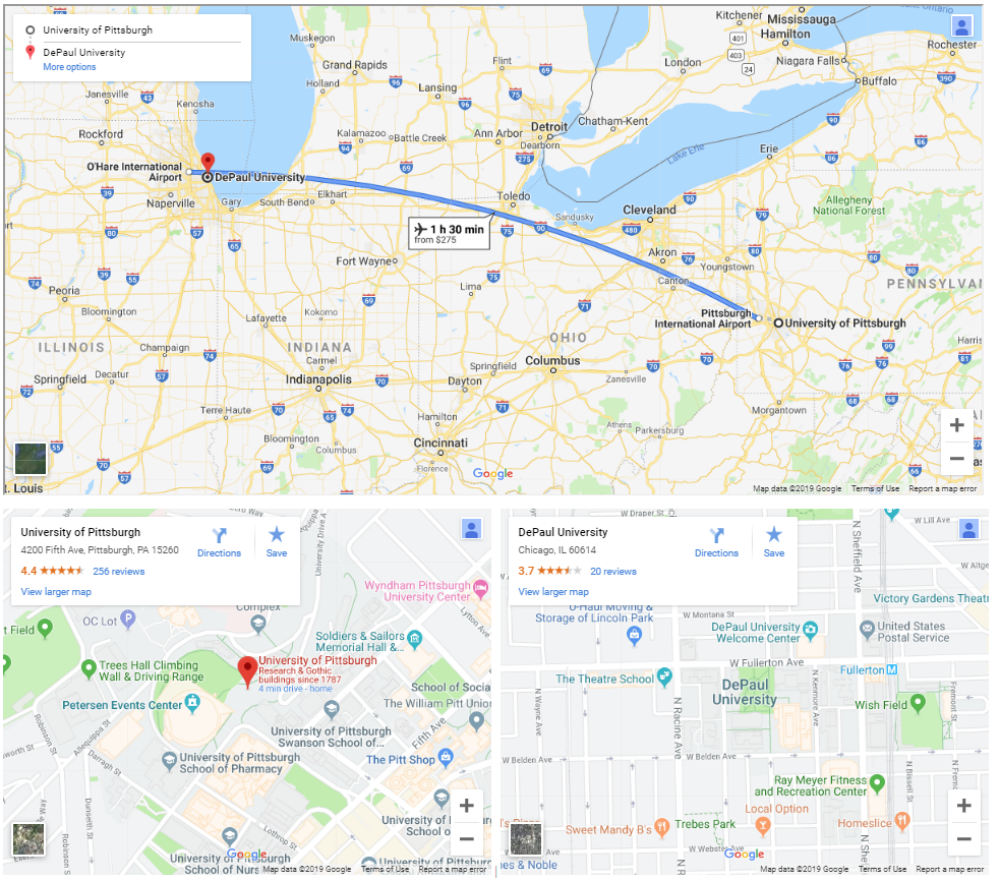


Figure 25: The visual interfaces that used to explain *geography* similarity-based recommendation model in study 3.

scale (1=Strongly Disagree/Very Low, and 5=Strongly Agree/Very High).

6.2.4.1 Behavior Difference User activity while performing the recommendation-sorting tasks, was logged. All explanation interfaces were static, so the user behavior was relatively simple. I tracked the number of *mouse clicks* (click to view explanation interface) as well as the *time spent* in each task. The result of the log analysis is reported in Table 11. To analyze behavior differences among treatments, I performed Wilcoxon Rank Sum and Signed Rank Test on log activity variables. The normality assumption did not hold in my analysis.

Table 11: Log activity analysis of study 3

	Clicks	Time (Secs)
	M (SE)	M (SE)
Sim1	11.16 (1.68)	383.27 (206.70)
Sim1+	**18.22 (5.74)****	383.94 (165.03)
Sim2	5.17 (0.39)	346.58 (122.23)
Sim2+	**10.37 (2.14)****	399.88 (132.56)
Sim3	6.11 (1.67)	406.72 (197.01)
Sim3+	6.66 (3.85)	450.22 (269.12)
Sim4	6.00 (1.57)	308.00 (176.43)
Sim4+	6.94 (2.38)	348.88 (172.77)
Sim5	7.00 (2.44)	357.22 (142.168)
Sim5+	6.77 (2.94)	373.94 (176.58)

- **Text similarly group:** there was a significant difference in the number of clicks for *Sim1* (M=11.16, SD=1.68) and *Sim1+* (M=18.22, SD=5.74) interface; W(18)=9.5, $p < 0.01$. The users clicked more, i.e., inspecting the explanation interface more, in the interface of *Sim1+*. I did not find a significant effect on the time spent, but the time variance of *Sim1+* was smaller.
- **Topic similarly group:** there was a significant difference in the number of clicks for *Sim2* (M=5.17, SD=0.39) and *Sim2+* (M=10.37, SD=2.14) interface; W(18)=0, $p < 0.01$. The users clicked more, i.e., inspecting the explanation interface more, in the interface of *Sim2+*. I did not find a significant difference for the time spent, but the subjects took a longer time at average to complete the sorting task while using the *Sim2+* interface.
- **Co-authorship similarity group:** I didn't find significant difference between *Sim3* and *Sim3+*, in both clicks and time variables. However, there was a similar tendency that the users tended to click more and spend more time to solve the given tasks.

- **Item similarly group:** I did not find significant differences for clicks or time spent, but I can observe that the enhanced interface (*Sim4+*) required on average slightly more clicks and time to complete the tasks.
- **Geography similarity group:** I didn't find a significant difference between the two interfaces (*Sim5* and *Sim5+*), in both clicks and time variables. Although the users spent more time on *Sim5+* interface, it required fewer clicks in completing the given tasks.

In general, I found adding additional visual component resulted in more clicks and time spent to complete the sorting tasks. The combined explanation interface produced more user interactions than a single explanation. Furthermore, the tasks were demanding to the subjects since they spent at average 5 to 6 minutes to complete the sorting. The subjects faced more difficulties while interacting with the *Sim1* interfaces, which took the longest time and the most clicks to complete the task.

6.2.4.2 Survey Difference The survey feedback was collected after performing each of the recommendation-sorting tasks. The subjects were asked to answer questions for nineteen explanation factors and four NASA-TLX questions. I summarized the factor questions into seven exploratory goals (shown in Table 9), e.g., the goal of *Transparency (TP)* was consisted by the average score of Q1, Q2, Q3, Q4, and Q5, etc. The results of task survey and NASA-TLX survey were reported in Table 12 and Table 13, respectively. To analyze behavior differences among treatments, I performed Wilcoxon Rank Sum and Signed Rank Test on log activity variables. The normality assumption did not hold in my analysis.

- **Text similarly group:** the enhanced interface (*Sim1+*) received significantly higher ratings in the goal of *Transparency (TP)*; W(18)=97.5, $p < 0.05$, *Scrutability (SC)*; W(18)=54, $p < 0.01$, *Trust (TS)*; W(18)=96.5, $p < 0.05$, and *Effectiveness (ET)*; W(18)=73.5, $p < 0.01$. The result indicated that the baseline explanation interface (*Sim1*) benefited from the additional explanation component. I further analyzed the result of NASA-TLX survey and found similar effects. The subjects perceived significantly better performance (*TLX4*) in accomplishing the sorting task. I did not find significant

Table 12: Task survey analysis of study 3

	TR	SC	TS	PE	ET	EF	SA
	M (SE)	M (SE)	M (SE)	M (SE)	M (SE)	M (SE)	M (SE)
Sim1	3.37 (0.77)	3.22 (0.76)	3.33 (0.77)	3.62 (0.78)	3.25 (0.66)	3.41 (0.82)	3.40 (0.70)
Sim1+	**3.92 (0.71)***	**4.26 (0.77)****	**3.88 (0.64)***	3.87 (0.83)	**3.92 (0.64)****	3.66 (0.98)	3.62 (0.72)
Sim2	3.64 (0.91)	4.02 (0.81)	3.70 (0.92)	4.15 (0.60)	3.68 (0.85)	3.38 (0.92)	3.70 (0.69)
Sim2+	4.00 (0.78)	4.05 (0.96)	4.01 (0.75)	4.09 (0.66)	4.00 (0.80)	3.63 (0.92)	3.91 (0.66)
Sim3	3.66 (0.66)	3.56 (0.79)	3.70 (0.98)	3.38 (0.90)	3.59 (0.71)	3.16 (0.76)	3.37 (0.70)
Sim3+	3.91 (0.61)	3.79 (0.87)	4.01 (0.68)	3.53 (0.92)	4.03 (0.65)	3.58 (0.80)	3.50 (0.76)
Sim4	3.71 (0.83)	3.76 (0.80)	3.75 (0.88)	3.92 (0.87)	3.59 (1.04)	3.13 (0.70)	3.65 (0.61)
Sim4+	3.80 (0.74)	3.81 (1.04)	3.72 (0.85)	4.14 (0.77)	3.64 (0.93)	3.25 (0.89)	4.04 (0.71)
Sim5	3.15 (0.99)	3.23 (1.13)	3.22 (0.91)	3.24 (0.75)	3.33 (0.97)	2.66 (1.11)	3.04 (0.90)
Sim5+	3.05 (1.04)	3.38 (1.19)	3.18 (1.06)	3.03 (1.00)	3.33 (0.90)	2.77 (1.03)	3.08 (1.00)

differences in other questions, however, the users reported lower mental demand (*TLX1*), effort (*TLX5*) and frustration (*TLX6*) while interacting with the enhanced explanation interface (*Sim1+*).

- **Topic similarly group:** I did not find any significant differences in explanation goals.

Table 13: NASA-TLX survey analysis of study 3

	TLX1	TLX4	TLX5	TLX6
	M (SE)	M (SE)	M (SE)	M (SE)
Sim1	2.77 (1.16)	3.66 (0.76)	2.61 (1.09)	2.05 (1.16)
Sim1+	2.55 (1.38)	**4.22 (0.94)***	2.16 (1.38)	1.77 (1.06)
Sim2	2.00 (1.38)	4.52 (0.51)	1.52 (0.79)	1.23 (0.75)
Sim2+	2.33 (1.28)	4.22 (0.73)	1.88 (0.90)	1.50 (0.70)
Sim3	2.50 (0.98)	3.77 (0.87)	2.61 (1.24)	1.88 (0.90)
Sim3+	2.77 (1.06)	3.83 (0.70)	2.61 (1.03)	1.94 (1.05)
Sim4	2.22 (1.35)	4.27 (0.82)	1.66 (1.02)	1.50 (0.92)
Sim4+	2.22 (1.55)	4.44 (0.70)	1.88 (1.32)	1.16 (0.51)
Sim5	1.94 (1.21)	3.72 (1.27)	1.94 (1.21)	1.44 (0.92)
Sim5+	1.72 (1.17)	4.38 (1.03)	1.77 (1.06)	1.55 (1.19)

However, I can still observe the similar improving tendency when adding a visual component to the baseline interface. The subject's perception increased on average for almost all explanation goals, except *Persuasiveness (PE)*. The result hints that the additional explaining component might improve the explainability of the baseline interface (*Sim2*). Interesting, in the survey of mental questions, I found the enhanced interface led on average to higher mental demand (*TLX1*), lower performance (*TLX4*), higher efforts (*TLX5*) and higher frustration (*TLX5*), although none of the differences were significant.

- **Co-authorship similarity group:** I did not find significant differences in the seven explanation goals. However, I can still observe the similar improving tendency when adding a visual component to the baseline interface. The subject perception score increased on average for all explanation goals. The result hints that the additional explaining component (*the edge thickness*) might improve the explainability of the baseline interface (*Sim3*). In the survey of mention questions, I found the enhanced interface led on average to higher mental demand (*TLX1*), higher performance (*TLX4*), the same level of

efforts (*TLX5*) and higher frustration (*TLX5*), although none of the differences were significant.

- **Item similarly group:** I did not find any significant differences in the item similarly group but observed the same tendency of improved user perception when adding a visual component to the baseline interface. The subjects' average perception increased for almost all explanation goals, except the goal of Trust (TS). The result indicated that the explainability of the baseline interface (*Sim4*) could be improved by adding a paper or user list to the Venn Tag interface. The survey of mental questions provided an interesting finding. I found the subjects perceived comparable mental demand (*TLX1*), higher performance (*TLX4*), higher efforts (*TLX5*) and lower frustration (*TLX5*). That is, the subjects did not feel a higher mental demand, yet adding an extra list still made them perceive the sorting task as harder to accomplish.

- **Geography similarity group:** I did not find any significant differences in the geography similarly group but observed a different user perception pattern when adding a visual component to the baseline interface (*Sim5*). I found adding a geo-location label may decreased the user perception score in the aspect of *Transparency (TR)*, *Trust (TS)*, and *Persuasiveness (PE)*. The result indicated that the explainability of the baseline interface (*Sim4*) may not always be benefited from the additional explanation component. In the survey of mention questions, I found the enhanced interface led on average to higher mental demand (*TLX1*), higher performance (*TLX4*), lower efforts (*TLX5*) and higher frustration (*TLX5*), although none of the differences were significant.

In general, I found that adding a visual component leads to a higher user perception score in the explanation goals. However, the improvement in the explanation goals did not guarantee a better user mental model. I found the interface *Sim1* was benefited the most by the additional visual component, either in user perception or mental survey. Adding a visual component to *Sim2* improved user perception but impaired the user mental model. In the interface *Sim4*, adding the extra component improved the user perception while maintaining a comparable user mental model.

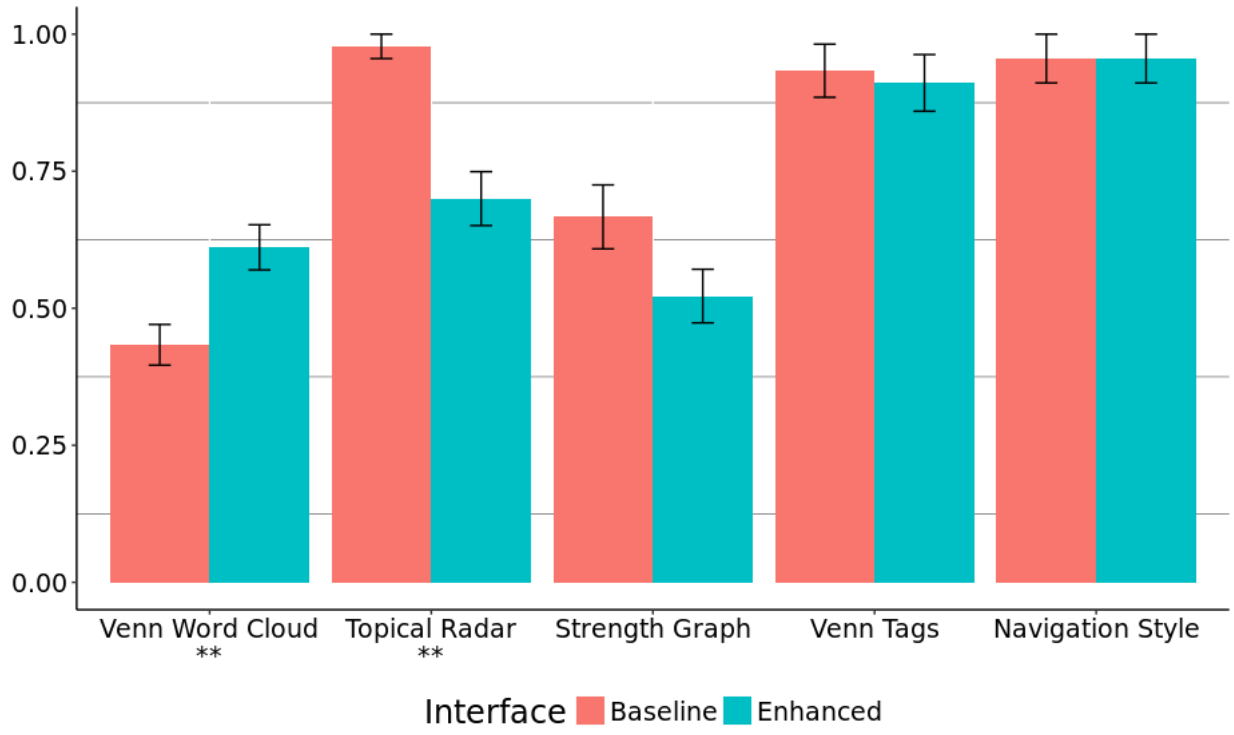


Figure 26: The correct rate of the recommendation-sorting tasks of study 3. Statistical significance level: (**) $p < 0.01$; (*) $p < 0.05$)

6.2.4.3 Sorting Difficulty In addition to the subjective feedback, I am interested in the question of how do the interfaces help the user to compare (sort) the relevance across recommendations. In each interface, I generated five recommendations using the associated recommendation model with a sample scholar profile. I then asked the subjects to sort the relevance among the five given recommendations and compared the answer with the ground truth. I used the *correct rate* to define the *sorting difficulty* among the explanation interfaces. It was an essential metric of *performance* when the user adopted the explanations interfaces in the exploring recommendations. The result was reported in Figure 26.

- **Text similarly group:** there was a significant difference in the correct rate for *Sim1* (M=0.43, SD=0.15) and *Sim1+* (M=0.61, SD=0.17) interface; W(18)=75.5, $p < 0.01$. The result was surprising to show the subjects achieve only 43% correct rate when at-

tempting to sort the given five recommendations with relevance. In this case, adding a visual component can be pretty helpful in assisting the subjects in completing the sorting task. However, a 61% correct rate may not be considered as an effective explanation interface, in particular, when the users have a chance to browse multiple recommendations and compare the explanations. The inconsistency would hurt the user trust and satisfaction to the explanation interfaces.

- **Topic similarly group:** there was a significant difference in the correct rate for *Sim2* (M=0.97, SD=0.09) and *Sim2+* (M=0.70, SD=0.20) interface; W(18)=286, $p < 0.001$. I found adding extra visual components impaired the judgment on sorting the recommendation relevance. In the baseline interface (*Sim2*), the subjects can achieve a 97% correct rate, which is strong evidence to support the explanation interface did help the users to sort the recommendation relevance. However, when adding the extra two topical word clouds, I found the correct rate was significantly decreased to 70%, which indicated the users might be "mislead" by the extra information. The result implied that adding the extra visual component can misinform the user, although the explanation interface was preferred and received higher user perception ratings by the user.
- **Co-authorship similarity group:** there was no significant difference in the correct rate for *Sim3* (M=0.66, SD=0.24) and *Sim3+* (M=0.52, SD=0.20) interface, the correct rate is between 52% to 66%. The result represents the graph-based explanation interface that may mislead the users in comparing the recommendation relevance, which calculated by scholar co-authorship networks. Adding an extra component made it become more difficult to compare tasks. Due to this, the two explanation interfaces may not be considered as an effective explanation interface.
- **Item similarly group:** I did not find a significant difference in the correct rate, in the item similarly group. Both of the interfaces helped the user to achieve a high correct rate (¿90%): *Sim4* (M=0.93, SD=0.20) and *Sim4+* (M=0.91, SD=0.21). The result implied adding an extra list to the Venn Tag diagram may not impair or improve the user inspectability (performance) of sorting the recommendations.
- **Geography similarity group:** I did not find a significant difference in the correct rate, in the geography similarly group. Both of the interfaces helped the user to achieve a high

correct rate (> 90%): *Sim4* (M=0.95, SD=0.18) and *Sim4+* (M=0.95, SD=0.18). The result implied adding an extra list to the map diagram may not impair or improve the user inspectability (performance) of sorting the recommendations.

6.2.5 Relations Between Survey and Log Variables

To better understand the relationship between the survey, log activities, and sorting result. I aggregated the variables in all three tasks (N=54). I then performed a correlation (using $Pearson's\ r$) analysis between task survey items and log variables revealed some interesting associations. The result was reported in Table 14. In general, when subjects did more mouse click activities, the recommendation-sorting correct rate was decreased (*Correct Rate*, r=-0.44, $p < 0.01$) and the subjects will feel more frustrated (*TLX6*, r=0.20, $p < 0.05$). The mouse click means spent more time (*Time*, 0.15, $p = 0.12$) in completing the tasks. The longer time of completion negatively correlated to all explanation goals, e.g., lower the user perception in system transparency (*Transparency*, r= -0.21, $p < 0.05$).

The better inspectability means the subjects can correctly sort the recommendation by relevance. I found the subjects can better understand (*Scrutability*, r=0.21, $p < 0.05$), be convinced by (*Persuasiveness*, r=0.20, $p < 0.05$) and be satisfied (*Satisfaction*, r=0.25, $p < 0.01$) the explanation interface more when they can achieve high correct rate of recommendation-sorting task. Furthermore, the subjects tended to feel less mental demand (*TLX1*, r=-0.22, $p < 0.05$), less effort (*TLX5*, r=-0.39, $p < 0.01$), less frustration (*TLX6*, r=-0.24, $p < 0.01$) but feel more confident in answering the sorting question (*TLX4*, r=0.43, $p < 0.01$).

I also found high internal consistency among all seven explanation goals, which implied the post-experiment survey was reliable. The goal of transparency, trust, and effectiveness were highly correlated with each other, which was reasonable because they shared one common factor (the Q5 in Table 9). That is, the correlation analysis suggested that if I can provide an explanation interface with a high transparency rating, then I can assume the user may tend to trust and feel the effects in the recommendations.

Higher user perception in the goal of scrutability (r=-0.27, $p < 0.01$), trust (r=-0.23, $p < 0.05$), persuasiveness (r=-0.40, $p < 0.01$), and effectiveness (r=-0.29, $p < 0.01$) can

reduce the storing difficulty (*TLX5*). Since the mental variable of *TLX5* and *TLX6* were highly correlated with each other (r=0.65, $p < 0.01$), it was reasonable to expect if one explanation goal was negatively correlated with *TLX5* then it should matintain the same pattern with *TLX6*, e.g., between *Effectiveness* and *TLX6*. However, in the explanation goal of *persuasiveness*, I found a positive correlation with *TLX6* (r=0.35, $p < 0.01$), i.e., when the recommendations were very persuasive, the user tend to frustrate more in completing the sorting tasks. I believe this is due to the explanation interface required the users to inspect more details (i.e., the Q14 in Table 9), which leaded to a higher cognitive load.

6.3 SUMMARY AND DISCUSSION

In this section, I presented two user studies of explanation interfaces for three similarity-based recommendation models. In study 2, I compared 25 explanation interfaces (20 visual explanations and five text-based explanations) through nineteen explanation factors. The experiment results suggested that participants preferred visual explanation interfaces over text-based explanation interface. I selected top-rated interfaces to explain the recommendation model. Based on the post-stage user interview, I further proposed *enhanced* visual component to each explanation interface.

In study 3, I conducted a performance-focused evaluation of ten explanation interfaces. For each model, I compared the top-rated design (**baseline**) with a combination of top and second-rated interfaces (**enhanced**). I expected that the complementary nature of the top designs could make their combination even stronger than the top choice alone. I found, however, that adding another visual component may result in increasing the cognitive overload and even creating a mental conflict. The findings were varied of each recommendation model: in the group of text similarity, I found adding a new visual component (*Two-way Bar Chart*) to the original explanation interfaces significantly improves user performance. However, in the group of topic similarity, I found that adding a new visual component (*Word Clouds*) might impair the user perception and performance of the recommendation-sorting task. In the group of item similarity, the extra explanation (list) did not change the user

perception or performance scores.

Based on the outcome of two user studies, I found the proposed explanation interfaces did reach the explanation goals. The result of the task survey suggested that adding a visual component (enhanced explanation interface) might contribute to a higher user perception score in the explanation goals. However, the improved explanation goals did not guarantee a better user mental model, based on the index of NASA-TLX. The result of recommendation-sorting tasks further suggested the inspectability (performance) can be improved by adding the extra visual component, but the user-preferred interface may not guarantee the same level of performance. Finally, I introduced a correlation analysis to discuss the relationships between survey and user behavioral variables.

There are several limitations of the presented work. First, the scale of the conducted studies was small. Larger-scale studies are needed for more definitive conclusions. Second, user rating and post-stage question ordering are not normalized to control the potential bias. Third, I do not consider the user personality that may influence user interaction. Fourth, the recommendations were generated for the same sample system user rather than for subjects themselves. All these issues will be addressed in my future work through a larger-scale, lab controlled study.

Table 14: Correlation analysis of study 3 (N=54)

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1. Correct Rate	1.00													
2. Clicks	**-0.44** **	1.00												
3. Time	-0.04	0.15	1.00											
4. Transparency	0.10	0.03	**-0.21** *	1.00										
5. Scrutability	**0.21** *	0.06	-0.03	**0.58** **	1.00									
6. Trust	0.14	-0.01	-0.11	**0.83** **	**0.67** **	1.00								
7. Persuasiveness	**0.20** *	-0.09	0.11	**0.49** **	**0.39** **	**0.58** **	1.00							
8. Effectiveness	0.07	0.02	-0.09	**0.77** **	**0.58** **	**0.93** **	**0.58** **	1.00						
9. Efficiency	-0.03	0.15	-0.12	**0.47** **	**0.30** **	**0.51** **	**0.31** **	**0.48** **	1.00					
10. Satisfaction	**0.25** **	-0.10	-0.13	**0.65** **	**0.45** **	**0.59** **	**0.68** **	**0.51** **	**0.43** **	1.00				
11. TLX1	**-0.22** *	0.07	-0.14	-0.04	-0.10	-0.09	**-0.29** **	-0.17	0.00	-0.05	1.00			
12. TLX4	**0.43** **	-0.06	-0.05	0.13	**0.26** **	0.16	**0.41** **	0.17	0.00	0.17	**-0.49** **	1.00		
13. TLX5	**-0.39** **	0.15	0.06	-0.14	**-0.27** **	**-0.23** *	**-0.40** **	**-0.29** **	-0.01	-0.15	**0.66** **	**-0.72** **	1.00	
14. TLX6	**-0.24** *	**0.20***	0.15	-0.05	-0.11	-0.16	**0.35** **	**-0.24** *	-0.02	-0.13	**0.44** **	**-0.56** **	**0.65** **	1.00

7.0 EXPLAINING SOCIAL RECOMMENDATIONS IN AN INTERACTIVE HYBRID SOCIAL RECOMMENDER

In this chapter, I investigated the effects of adding *explanations* to an interactive hybrid social recommender system through study 4. I conducted an online user study (N=33) at three research conferences to evaluate user behavior and obtained subjective feedback of the six proposed explainable interfaces. This study intends to answer two research questions: 1) Which visualization is better for explaining an interactive hybrid social recommender system? 2) How do the explanations affect user perception and interaction with an interactive hybrid social recommender system? The findings can be summarized in three-fold: First, I confirm the user-driven fusion principle using a state-of-the-art user-controllable interface. Second, I provide a new exploratory model (with six explainable interfaces) for explaining an interactive hybrid social recommender system. Third, I show evidence to support the interaction effect between the factors of controllability and explainability.

7.1 INTRODUCTION

In study 4, I propose *relevance Tuner+*, an extension of my earlier system [132], which provides a controllable interface for the user to fuse social recommendations from multiple sources. A total of five recommender engines were introduced in this study:

1. *Publication Similarity:* cosine similarity of users' publication text.
2. *Topic Similarity:* topic modeling similarity of research interests (topics).
3. *Co-Authorship Similarity:* the network distance based on co-authorship networks.

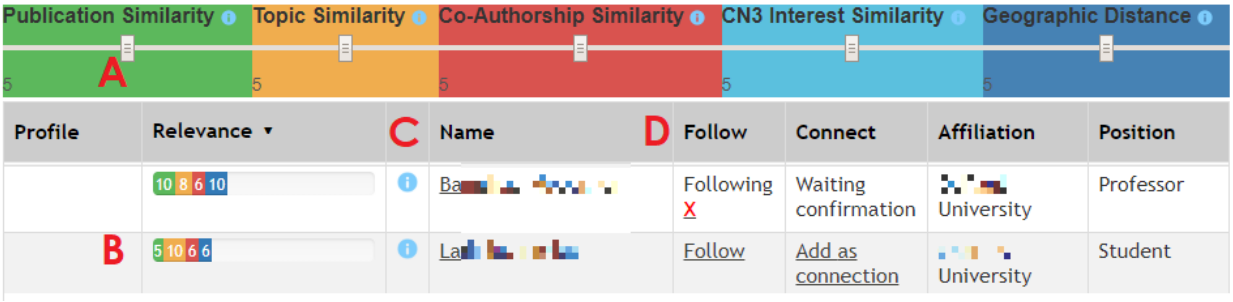


Figure 27: The design of Relevance Tuner+: (A) relevance sliders; (B) stackable score bar; (C) explanation icon; (D) user profiles.

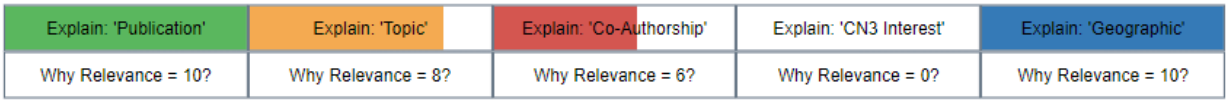


Figure 28: The pop-up window of clicking the *explanation icon* (shown in Figure 27 Section C).

4. *CN3 Interest Similarity:* the number of papers co-bookmarked, as well as the authors co-followed.

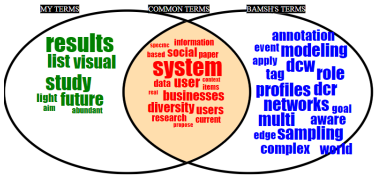
5. *Geographic Distance:* a measurement of the geographic distance between affiliations.

The users can "tune" (re-rank) the social recommendations using five sliders (see Figure 27, Section A). The user can *explore* the relevance scores (sum of personalized relevance score of five recommendation engines) through the colored stackable bars in Section B and *access* more information about recommended people using links in Section D. In this study, I introduced a new *explanation icon* in Section C. The user can *inspect* the relevance by clicking the icon. A window will pop up (Figure 28) to show a clickable explanation table. A click on the first-row cell will open a visual explanation of calculated relevance for the selected recommendation engine (as shown in Figure 29 (a) to (e)). A click on the second-row cell will show the calculation process of the relevance scores (Figure 29f). In this design, I attempted

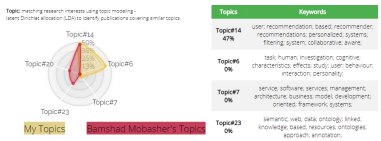
to separate the explanation for the fusion process, which the user can influence by tuning the sliders from the explanation of each relevance obtained by clicking the explanation icon. To be more specific, I intend to help the user to get explanations for two kinds of questions: fused relevance questions (i.e., *"Why this recommendation is ranked at the top?"*) and engine relevance questions (i.e., *"Why topic relevance is equal to 8?"*).

7.2 PRESENTATION OF EXPLANATIONS

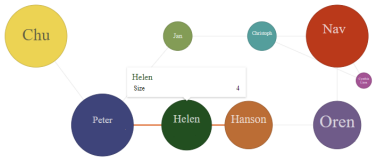
(a) Venn Word Clouds



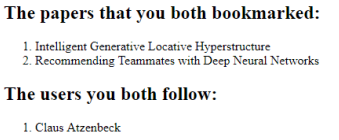
(b) Topic Similarity



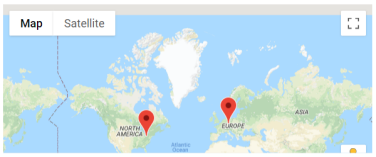
(c) Co-Authorship Graph



(d) Interest Similarity



(e) Geographic Distance



(f) Relevance Equation

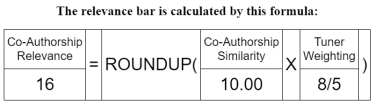


Figure 29: I presented six interfaces for explaining the social recommendations in study 4. a) Venn Word Clouds; b) Topic Similarity; c) Co-Authorship Graph; d) Interest Similarity; e) Geographic Distance; f) Relevance Equation. The visualization (a) to (e) is matched with the corresponding recommendation engine (first row of Figure 28), but (f) was adopted across five recommendation engines (second row of Figure 28).

Instead of using a context-specific visualized recommendations [70, 140], I added an *explanation icon* next to each social recommendation, leaving the choice of requesting the details behind the reasoning to the users. The information was provided by a hybrid explanation approach [103, 75], which mixed multiple visualization components for explaining the details of the recommendation engine. A total of five visual explanations and five equations

(one for each relevance) were proposed. I referred *the user* to present the current log-in user who is using the interface. I used *attendee* to represent the recommended conference attendee being inspected.

Publication Similarity: I adopted a *Venn word cloud* visualization inspired by *tag cloud* [41, 104] as an approach to explaining the text-level similarity between the publication of the user and the attendee (Figure 29a). This visualization presented the *terms* of the paper title and abstract. The font size indicates the term frequency in the documents. The user's terms and the attendee's terms will be presented on the left and right, respectively. The terms in the middle presented the *words-in-common*, which means the terms were appearing in the publications of both scholars.

Topic Similarity: I presented research topics in a radar chart and the topical words of each research topic in table [148]. The visualization design can be found in Figure 29b. The radar chart was presented on the left side. I selected the top 5 (ranked by *beta* value from a total of 30 topics) topics of the user and compared them with the attendee. A table with topical words was presented in the right so that the user can inspect the context of each research topic.

Co-Authorship Similarity: I presented co-authorship network in a path graph [131]. The visualization design can be found in Figure 29c. For connecting the user (yellow circle in the left) to the attendee (red color in the right), I tried to find six possible paths (one shortest and five alternatives) by direct and in-direct co-authorship. The circle size represented the connectivity, i.e., *Peter* is the only node that scholar *Chu* can reach scholar *Nav*, so the circle size was the largest one (size = 6).

Interest Similarity: I presented co-bookmarking (conference paper) / co-following (conference attendees) information in an itemized list, inspired by the user-based approach [103]. The visualization design can be found in Figure 29d. I used two itemized lists to show this information. The design helps the user to inspect the overlapped items that the recommendation engine used to calculate the similarity.

Geographic Distance: I plotted cities of affiliations on a world map, inspired by location-based explanation [103]. The visualization design can be found in Figure 29e. I bundled the *Google Map API* for presenting the geo-location of the two affiliations on the

map. The two *pins* were the affiliated institution of the user and the target user so that the user can inspect the geo-distance, regions, or the country information.

Relevance Equation: I used *relevance equation* (Figure 29f) to explain the calculation process of each of the five relevance scores shown in the stackable bars (Figure 27 Section B). The relevance score equals the recommendation engine similarity multiply by tuner weighting with a roundup function. For example, if the user tunes the *Co-Authorship Similarity* to 8 (Figure 27, Section A), then the normalized tuner weight is $\frac{8}{5} = 1.6$. As explained in Figure 29f, to obtain the resulting source relevance score 16, the tuner weight is multiplied by the source similarity score 10.

7.3 EXPERIMENT DESIGN AND PROCEDURE (STUDY 4)

The recommendations produced by all five engines are based on data collected by the Conference Navigator 3 system [10, 132]. To recruit the user study participants, I sent out invitation emails to attendees of three conferences. A total of 345 emails were sent. I received 43 responses (response rate=12.46%). After sending the personalized study link to all respondents, there were 33 participants (12 female) who eventually accepted and completed the online user study. Participation is voluntary. The participants attended Hypertext 2018 (10 participants); UMAP 2018 (12 participants), or EC-TEL 2018 (11 participants). Their ages ranged from 20 to 59 (M=31.00, SE=7.74). All of them could be considered as knowledgeable in their research area and had at least one academic publication at the corresponding conference.

To assess the value of the proposed interface, I compared the explainable and controllable Relevance Tuner+ interface (*Tuner+*) with a controllable-only interface (Section C in Figure 27 removed) (*Baseline*). The online study adopted a within-subjects design. A two-minute tutorial video was provided for participants to familiarize themselves with the interface before each treatment. All participants were asked to use each interface for three information-seeking tasks and to fill out a post-stage questionnaire. The order of question was the same to all participants with a 5-point scale (1=Strongly Disagree and 5=Strongly Agree).

The order of treatment was randomized to control for the effect of ordering (half of the participants started the study with the baseline interface). To minimize the learning effect (becoming familiar with the conference data), I used data from different years of the same conference (EC-TEL 2017 & 2018) or alternative conferences (HT/UMAP 2018) in the two treatments.

Participants were given the same three tasks for each treatment. The tasks were explicitly designed to be diverse but realistic tasks that could be naturally pursued by attendees of research conferences.

- **Task 1:** "Please use the system to follow four conference attendees you would like to talk during the coffee break."
- **Task 2:** "Your advisor asks you to follow four conference attendees with close connection with your research group. He/she would also appreciate that the scholars be from different regions of the world."
- **Task 3:** "Please use the system to find four committee member candidates for your dissertation defense. The candidates should be senior scholars with expertise close to your research field". The participants were asked to pick suitable candidates among conference attendees, based on their best judgment in each task.

7.4 DATA AND MEASUREMENTS

I collected action logs for slider manipulations, explanation clicks, and the time to complete the tasks the post-stage survey comprised of 16 questions that covered different user experience (UX) dimensions. In the *Tuner+* treatment, three extra questions were presented for collecting the user feedback on each explanation design. I then built a structural equation model (SEM) for analyzing the UX concepts and the directionality of causal effects. I followed the framework introduced in [71]. I planned three latent constructs: two subjective system aspects (SSA) (perceived control & perceived transparency) and one user experience (EXP) (satisfaction). The model fit the statistics of $\chi^2(66) = 411.65$, $p < 0.001$,

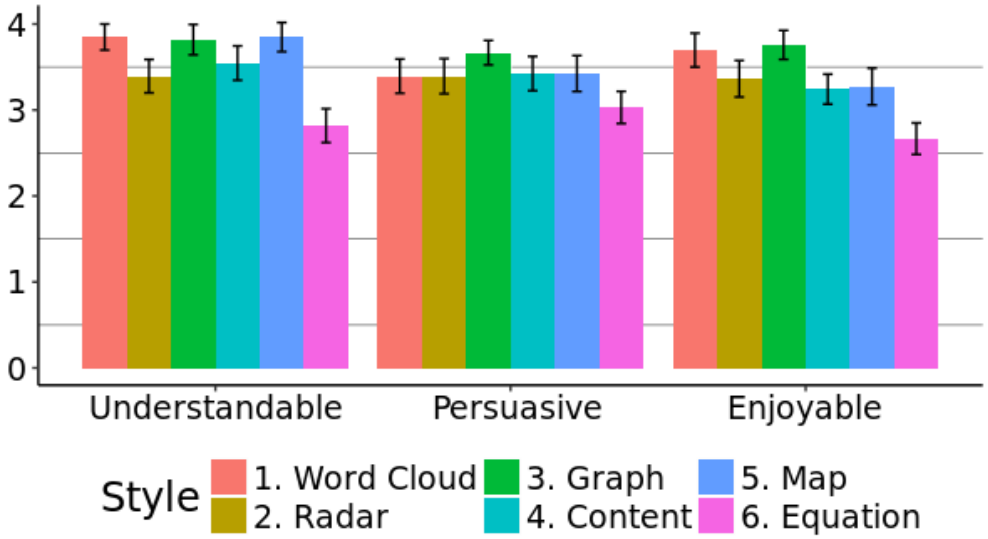


Figure 30: The user feedback of six explanation styles of study 4.

$RMSEA = 0.18$, $90\%CI : [0.13, 0.17]$, $CFI = 0.88$, $TLI = 0.89$. The three factors listed below showed good convergent validity (AVE) and internal consistency (Cronbach's α).

- *Perceived Control* (SSA) ($AVE = .60, \alpha = .77$): 3 items, e.g., "I feel in control of modifying my preferences","I became familiar with the recommender system very quickly".
- *Perceived Transparency* (SSA) ($AVE = .55, \alpha = .81$): 5 items, e.g., "The recommender explains why the conference attendees are recommended to me","I understood why the contacts were recommended to me".
- *Satisfaction* (EXP) ($AVE = .64, \alpha = .91$): 8 items, e.g., "The recommender helped me find the ideal contacts at conference", "Overall, I am satisfied with the recommender".

7.5 RESULTS

In the *baseline* group, I found that the participants extensively used the relevance sliders to complete the three assigned tasks ($M = 56.78$, $SD = 42.21$, $User\ Count = 30$). There

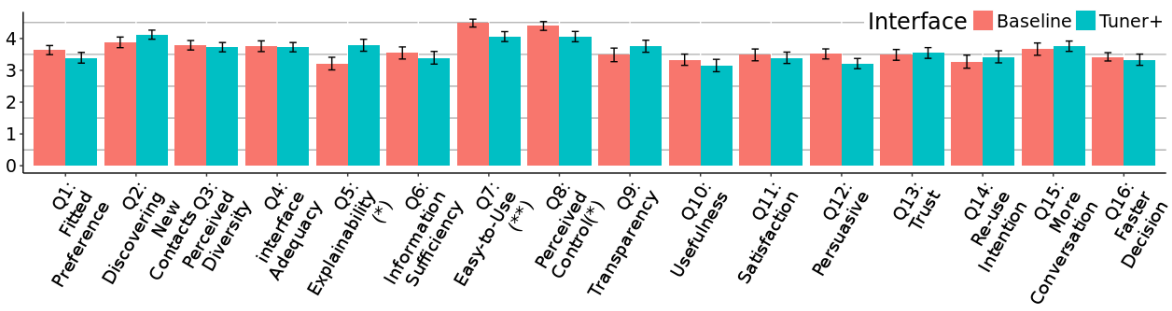


Figure 31: User feedback analysis of study 4: the result shows that the *Tuner+* interface received a significantly higher rating in the aspect of explainability (Q5). (Statistical significance level: (**) $p < 0.01$; (*) $p < 0.05$)

were only three users who didn't interact with the sliders. In the *Tuner+* group, I found a similar pattern in using the sliders ($M = 64.63$, $SD = 43.84$, $User\ Count = 32$) that almost all participants did interact with the sliders. However, only around 50% of the participants clicked on the explanations icon ($M = 5.90$, $SD = 8.36$, $User\ Count = 17$), i.e., the relevance sliders were used by the participants more extensively than the explanation icon. The participants spent more time to complete the three assigned tasks in *Tuner+* group i.e., when the explanations were provided ($M = 733.69$, $SD = 766.10$, in seconds), than *baseline* group ($M = 573.21$, $SD = 567.48$).

Figure 30 shows user feedback on the six explanation styles. For a more *understandable* visualization, the explanation style of *word cloud*, *graph*, and *map* received higher scores. For better *persuasive*, i.e., convincing the user to accept the recommendation, the explanation style of *graph* outperformed the other visualizations. One participant specifically commented that the social network visualization is *"really interesting and useful"*. For better satisfying the user (*enjoyable*), the explanation styles of *word cloud* and *graph* were preferred by the participants after experiencing the system.

I performed a Wilcoxon signed-rank test for analyzing the subjective feedback (shown in Figure 31). I found that many user-ratings are comparable between treatments, which means that adding an explanation icon to the system does not impact the UX dimensions. However,

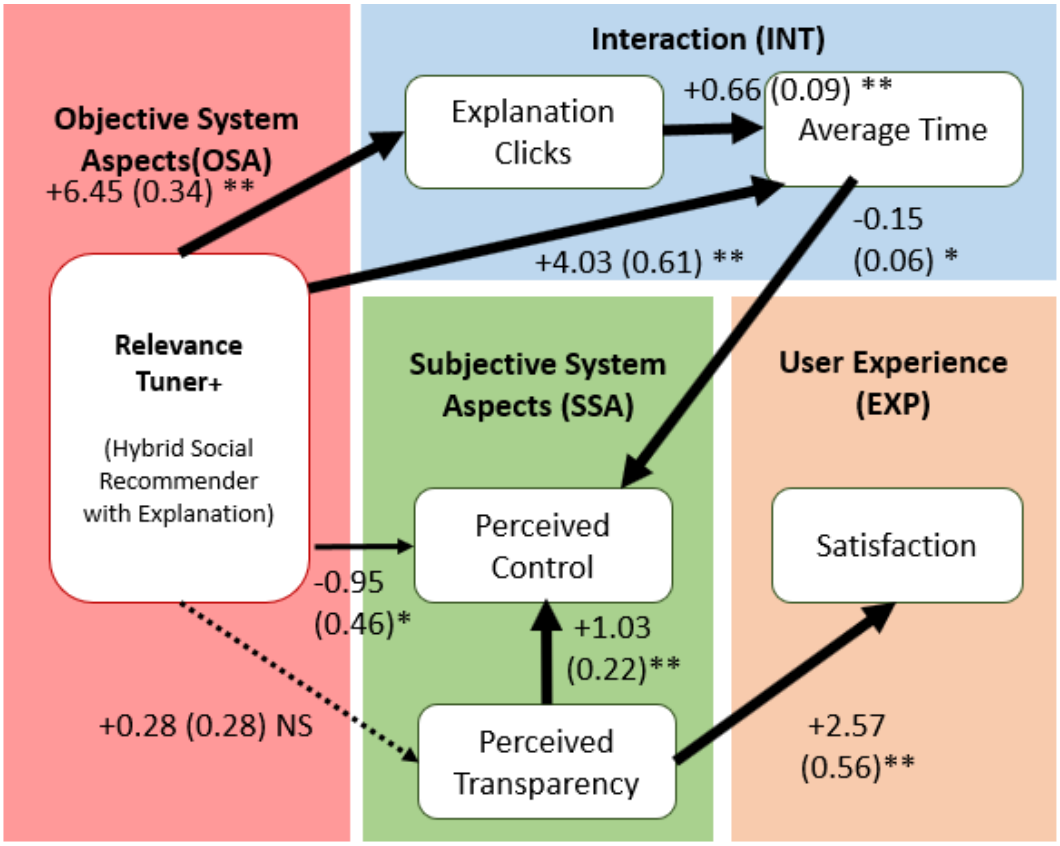


Figure 32: The structural equation model (SEM) of study 4. The number (thickness) on the arrows represents the β coefficients and standard error of the effect. (Statistical significance level: (**) $p < 0.01$, (*) $p < 0.5$, (NS) no significance)

I found that the participants agreed the *Tuner+* interface significantly better in providing *explainability (Q5)*, which indicated the attached explanations were useful in gaining system transparency and providing the reasoning process of social recommendations. Interestingly, I also found that if the explanations were presented, the participants' perception of the *easy-to-use interface (Q7)* and *perceived control (Q8)* were significantly decreased, which implied the users might experience difficulties with a possibly overwhelming amount of information.

I confirmed the finding in SEM analysis (shown in Figure 32). I found that adding extra explanations (OSA) decreases the user perception of controllability (SSA). In *Tuner+* condition, the participants to click the explanation icons more (INT) to inspect the social

recommendations, which increases the average time spent (INT) in completing the tasks. If more time spent on each task, the subjective system aspect (SSA) on *perceived control* decreases. However, I also found the participants who perceived more *transparency* will positively associate this with *perceived control* and *satisfaction*. The pattern implied the extra amount of information would not impair those who perceived higher system explainability and understanding.

7.6 SUMMARY AND DISCUSSION

This section presents an evaluation of explainable recommendations in an interactive hybrid social recommender, *Relevance Tuner+*. My works extended the earlier version of the controllable interface *Relevance Tuner* with explanations in the form of five visualizations and five equations. I found that the user extensively uses sliders to adjust source weights while completing the conference-attendee exploration tasks. The result supports prior findings [132] that an interactive interface helps to improve the user experience and initiate user-driven exploration. At the same time, the explanations were not used as heavily. Among visual explanations *word cloud* and *graph* were rated with a higher score in the aspects of understandable, persuasive and enjoyable (shown in Figure 30).

I also found an interesting perception trade-off between controllability and explainability. More specifically, the experiment result indicates that when users can inspect the social recommendation with an on-demand explanation, it increases their perception of system explainability. However, the improvement comes with a price of reducing the user perception of control (Q8) and the sense of ease of use (Q7). I confirmed this finding in the analysis of SEM that shows the time spent in inspecting the social recommendations was negatively correlated with the factor of *Perceived Control*, a possibly overwhelming amount of information caused the users to decrease the perception of controllability. Although I didn't find a direct effect between providing an explaining icon and the user perception of transparency, it nonetheless plays a crucial role in contributing to the factor of user satisfaction.

The finding of controllability and explainability trade-off is surprising, but not an un-

charted area in the field of HCI. In explaining recommendations, one main goal is to provide completeness of information so the users can gradually improve the mental model while interacting with the system [77]. However, a detailed, full explanations may be "excessive" to the users [84], which had a negative impact on user confidence and enjoyment [114]. To overcome this problem, based on the context of information-seeking tasks, only the filtered "relevant, and important information" should be presented as explanations [32]. In this section, I asked the study participants to find conference attendees in different scenarios, e.g., "with close connection with your research group". In this case, the essential information is those recommendations with high "Co-Authorship Similarity", which can be done easier with the controllable sliders. The additional explanations may be attractive but not mandatory. When the overwhelming amount of information was provided, especially for those who didn't adopt the explanation interfaces, it impaired the user perception of controllability.

My work has some limitations. First, it is a small-scale user study (N=33). Second, in a semi-controlled online study, I found only half of the subjects explored the explanation functions (manipulated aspect), which may hurt the significant effect on the transparency factor in my SEM analysis. Third, user rating and post-stage question ordering may be biased by the rating tendency of each subject and might be better to normalize them. Fourth, there are too many variables in my experimental design. Further investigation will be required to control the interaction effects. All these issues will be addressed in my future work with a larger-scale, lab controlled study to confirm the findings and model robustness.

8.0 EVALUATING EXPLANATION INTERFACES USING CROWDSOURCING APPROACH

In this chapter, I performed a card-sorting analysis to identify the user preferred interfaces for explaining the five recommendation models. I found the interface design of *E1-4: Venn Word Cloud*, and the participants favored *E3-4: Strength Graph* in study 2. I then conducted the study 3 for evaluating the top-rated and second-rated interfaces. However, based on the experiment result of study 3, I found two explanation interfaces, *publication similarity* and *co-authorship similarity*, did not effectively support the sorting task of recommendation relevance. In this chapter, I aim to improve the explanation interface designs by conducting the second round evaluation, i.e., study 5.

8.1 INTRODUCTION

In the analysis of *Sim1 & Sim1+*, I found adding a visual component can assist the subjects in completing the sorting task. That is, the user performance in the group of *E1-4 Venn Word Cloud* could be improved by the additional term bar chart. Based on the after study user feedback, the study participants pointed out the original word cloud was hard for them to complete the comparing task. The style of word cloud was good for having an overview of a group of text, but it is hard to tell the difference between the two word clouds. In contrast, the participants also mentioned the bar chart would be clear and more straightforward for them to complete the comparing tasks. Since the user performance is the focus on the explanation interface, I decide to change the design to *Two Way Bar Chart*.

In explaining a text-based recommendation using the *Two Way Bar Chart*, I need to

determine two settings. First, it is required to determine the *number of terms* that presents in the bar chart. The ideal case is to show all terms in one figure. However, in the recommendation model, the length of the term vector is around 45,000, which makes it is not realistic to show all the terms in one interface. Hence, there is a need to conduct an experiment to determine the number of terms to explain the text-based recommendation model better. I choose two conditions as 30 terms and 60 terms. Second, a typical bar chart was ordered by *one dimension relevance*. However, in the *Two Way Bar Chart*, the relevance can be displayed by two dimensions. I propose two order methods here: *order by individual relevance* or *order by mutual relevance*. The method of *order by individual relevance* (shown in Figure 33(a)) treats the bi-directional bar chart as independent. That is, the bar chart will order by its own relevance, from high to low the benefit of this method to show two term relevance distributions. *order by mutual relevance* (shown in Figure 33(b)) will order the term mutually, i.e., sum of the bi-directional relevance and order them from high to low relevance. It is an approach to present mutually important terms on the top. It is easier for the user to perceive the *high impact* terms.

In the analysis of *Sim3* & *Sim3+*, I found both of the *Sim3* interfaces (*Strength Graph*) were performing better in user performance than the interface of *Sim3+*. The enhanced interface (*Sim3+*) highlighted the thickness of the network edges, i.e., the user performance did not improve by the additional edge thickness information. It is not clear here that if adding extra information can improve or impair the user performance. Hence, I would like to conduct a larger-scale user study to confirm the effective design.

In explaining a text-based recommendation using the *Strength Graph*, the graph was consisted of two components: edges and nodes. It is important to find out the most effective settings so the users can better understand the recommendation though the interface. First, it is important to define the *edge thickness* and determine the benefit of adding it to the *Strength Graph*. Usually, the edge thickness represents the *strength* between two connected nodes. In this case, it shows the strength between two scholars that I defined it as *the number of co-authored papers*. If more publication is co-authored by two scholars, the edge thickness will be increased. However, it is not clear if adding such information is a benefit to understand the co-authorship recommendation model. The design is shown in Figure 34(a)).

Second, the node size can also be controlled to show the *importance* of the node. In my case, I define the important as the *number of papers* that published by the scholar. It is a signal to show if one scholar's seniority in the network. However, it is also not clear if adding such information will help the user to understand the co-authorship recommendation model. The design is shown in Figure 34(b)).

8.2 EXPERIMENT DESIGN AND PROCEDURE (STUDY 5)

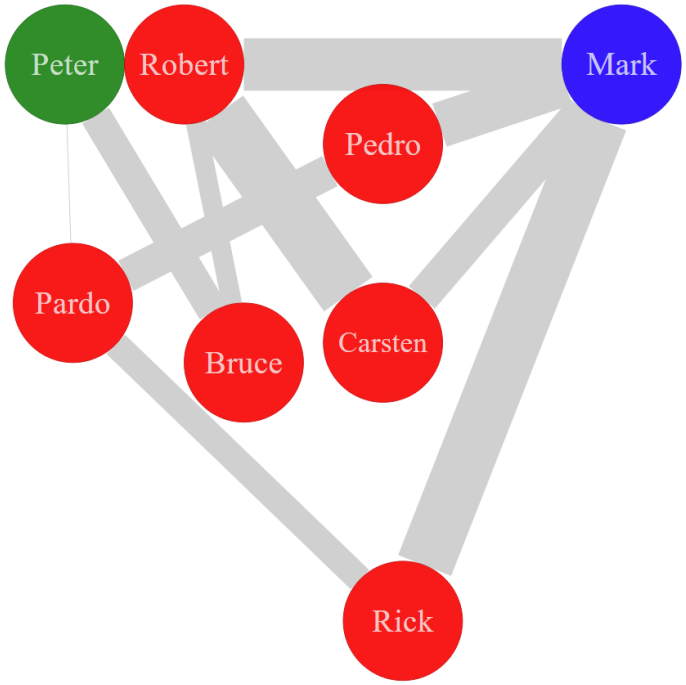
To evaluate the effectiveness of the two explanation interfaces, I conducted a randomized between-subject experiment on Amazon Mechanical Turk (MTurk) (study 5). We used a 2x2 design, resulting in four conditions of each interface. In the group of *Two Way Bar Chart*, there are four conditions: *Chart 1:* 30 terms with order by individual relevance, *Chart 2:* 60 terms with order by individual relevance, *Chart 3:* 30 terms with order by mutual relevance, and *Chart 4:* 60 terms with order by mutual relevance. In this group of *Strength Graph*, there are four conditions: *Graph 1:* disabled edge thickness and disabled node size, *Graph 2:* enabled edge thickness and disabled node size, *Graph 3:* disabled edge thickness and enabled node size, and *Graph 4:* enabled edge thickness and enabled node size. The participants were allowed to spend up to 24 hours to complete the study.

In study 5, I presented two figures (the same condition with high and low relevance) and a short instruction of the interface. To make the instruction can easily follow by participants from a diverse background. I chose to make the instruction from everyday life scenarios. In the group of *Two Way Bar Chart*, the participants were told to distinguishes two bar chat for different the similarity of *hash-tags* of two Twitter accounts. In the group of *Strength Graph*, the participants were asked to determine the probability of connecting friendship on social media, base on the two given social networks. The detail of the instructions was shown below.

- *Instruction of Two Way Bar Chart group:* "Hashtag" is a type of metadata tag used on social networks such as Twitter. "Hashtag" makes it possible for others to find messages with a specific theme or content easily. In this study, we present two pairs of users'

Figure 33: Explanation Interface of *Publication Similarity* using new Two-way Bar Chart: (a) Order by individual relevance, (b) Order by mutual relevance.

(a) Adding edge thickness



(b) Adding node size

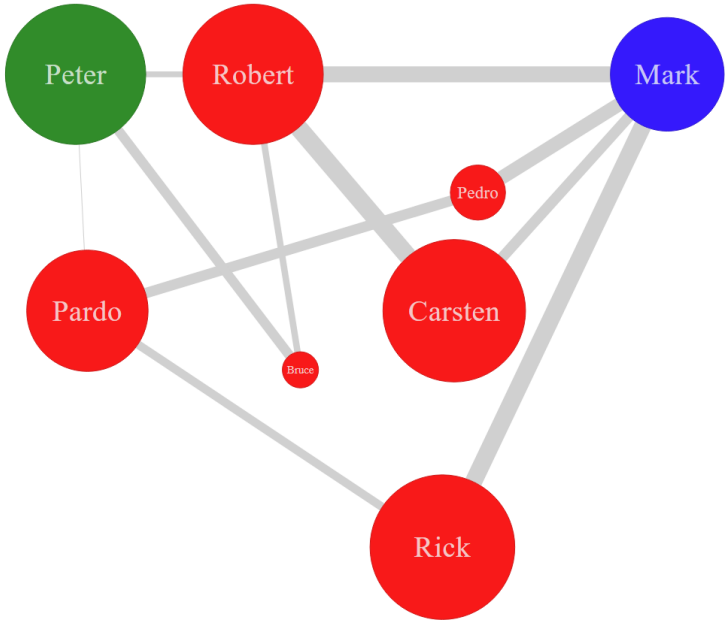


Figure 34: Explanation Interface of *Co-authorship Similarity* using Enhanced Strength Graph: (a) Adding edge thickness. (b) Adding node size.

Twitter Hashtags, Peter/Mary & Peter/Kelly, in two bar charts. In the bar chart, we present the name of hashtags on the y-axis and the "number of hashtags" in the x-axis. Please answer the questions based on the two bar charts.

- *Instruction of Strength Graph group:* Please answer the following questions based on the two social network visualizations. In both networks... Peter has several directed online contacts. The red bubbles represent a portion of the large network accessible through Peter's social media profile. The blue bubble is the person Peter wants to contact. The thickness of the edge represents the number of share friends (i.e., the thicker edge, the more shared friends between the two people) The node size represents the number of friends (the larger size, the more friends he/she has).

The subjects were expected to answer one testing question, one task question, and four NASA-TLX questions [46]. The testing questions were simply asked the subjects to find out information on the visualization [92, 20], i.e., *What is the top (most popular) hashtag in Peter's Twitter account?* and *What are the names labeled in blue color?*. The task question tests if the subject can tell the difference between the two figures, it is a multiple-choice question, shown below.

1. If we want to measure the "hashtag similairty", i.e., if two users shared more hashtags, then the "hashtag similairty" is higher. Which statement is true?
   - (Correct) Option A : "Peter & Mary"'s hashtag similarity is higher than "Peter & Kelly".
   - Option B: "Peter & Mary"'s hashtag similarity is lower than "Peter & Kelly".
   - Option C: The hashtag similarity of "Peter & Mary" and "Peter & Kelly" is the same.
   - Option D: The information is insufficient to determine the hashtag similarity.
2. If Peter wants to connect a new friend on his social media, which one of the following statements is correct?
   - (Correct) Option A: "Peter & Mark" is more likely to be connected than "Peter & Jones".
   - Option B: "Peter & Mark" is less likely to be connected than "Peter & Jones".

- Option C: The probability of connecting "Peter & Mark" is the same as "Peter & Jones".
- Option D: The information is insufficient to determine the probability of connecting.

The NASA-TLX question included: *(TLX1) Mental Demand: How mentally demanding was the task? (TLX4) Performance: How successful were you in accomplishing what you were asked to do? (TLX5) Effort: How hard did you have to work to accomplish your level of performance?* and *(TLX6) Frustration: How insecure, discouraged, irritated, stressed, and annoyed were you?* The order of question was the same to all participants with a 5-point scale (1=Strongly Disagree/Very Low, and 5=Strongly Agree/Very High).

The participants were randomly assigned to one of the four conditions. Each participant received a payment of USD $0.1 if their submission was accepted. In the group of *Two Way Bar Chart*, there were 458 participants joined the study and 400 participants passed the testing question. The participants was equally distributed to the *Chart 1* to *Chart 4*, i.e., 100 subjects for each condition. In the group of *Strength Graph*, I requited 435 participants and 417 participants passed the testing question. The participants were distributed as 111 subjects in *Graph 1*, 109 subjects in *Graph 2*, 103 subjects in *Graph 3*, and 103 subjects in *Graph 4*.

8.3 RESULTS

The condition of *Chart 2* (M=6.38, SD=8.23, in minutes) took the participants more time to complete the study. The condition of *Chart 1* (M=6.08, SD=8.21) took less time to complete. The condition of *Chart 2* (M=6.38, SD=8.23) and *Chart 3* (M=6.77, SD=10.23) was ranked in the middle. The subjects took surprisingly more time to complete the tasks, which indicates the visualization may not be east-to-use to the subjects. However, due to the mechanism of Amazon Murk, the participants can complete the task within 24 hours, the actual execution time should be shorter than the reported number. I report the percentage of each option in Table 15. I found the *Chart 2* has the highest correct rate at 65%, which indicates the higher number of terms and order by individual relevance was easier for the

user to compare the different between the two bar charts. I also find a pattern between the order methods. In the condition of ordered by mutual relevance (*Chart 3* and *Chart 4*), the participants took less time to complete the study but it did not guarantee a higher correct rate. The TLX analysis of the group of *Two Way Bar Chart* is reported in Table 17. The finding indicates the *Chart 2* has lower score of mental demand (TLX1) and feel of difficulty (TLX5). Moreover, the participants reported a higher score on perceiving success (TLX4) of solving the task. *Chart 2* also has a higher score of feel of stress (TLX6). I did not find any significance between the survey scores and spent time.

The condition of *Graph 2* (M=9.78, SD=10.99) took the participants more time to complete the study. The condition of *Graph 3* (M=6.23, SD=10.91) took less time to complete the study. The condition of *Chart 1* (M=7.90, SD=10.88) and *Chart 4* (M=7.04, SD=10.91) was ranked in the middle. The subjects took surprisingly more time to complete the tasks, which indicates the visualization may not be east-to-use to the subjects. However, due to the mechanism of Amazon Murk, the participants can complete the task within 24 hours, the actual execution time should be shorter than the reported number. I report the percentage of each option in Table 16. I found the *Graph 2* has the highest correct rate at 63%, which indicates enabled edge thickness, and disabled node size was easier for the user to compare the differences between the two strength graphs. I also find a pattern between the edge conditions. In the condition of enabled edge thickness (*Chart 1* and *Chart 3*), the participants took more time to complete the study, which also guarantees a higher correct rate. The TLX analysis of the group of *Strength Graph* is reported in Table **??**. The finding indicates the *Graph 2* has a higher score of mental demand (TLX1), perceiving success (TLX4), and a feel of difficulty (TLX5). The finding indicates the design of *Graph 2* is an effective design, but the users may need extra cognitive effort to interact with the interface. I did not find any significance between the survey scores and spent time.

Table 15: User Feedback Analysis of study 5: Two-way Bar Chart

Condition	Number of Terms	Order Method	Option A	Option B	Option C	Option D
Chart 1	30	Individual	61%	25%	11%	9%
Chart 2	60	Mutual	65%	25%	10%	4%
Chart 3	30	Mutual	50%	37%	11%	8%
Chart 4	60	Individual	56%	26%	14%	8%

Table 16: User Feedback Analysis of study 5: Enhanced Strength Graph

Condition	Edge Thickness	Node Size	Option A	Option B	Option C	Option D
Graph 1	Disabled	Disabled	60%	17%	18%	8%
Graph 2	Enabled	Disabled	63%	23%	21%	3%
Graph 3	Disabled	Enabled	57%	27%	13%	6%
Graph 4	Enabled	Enabled	59%	22%	16%	10%

Table 17: TLX analysis of study 5: Two-way Bar Chart

Condition	TLX1	TLX4	TLX5	TLX6
	M (SD)	M (SD)	M (SD)	M (SD)
Chart 1	4.03 (1.57)	5.04 (1.50)	4.27 (1.77)	3.14 (1.70)
Chart 2	3.96 (1.67)	5.37 (1.44)	4.17 (1.71)	3.31 (1.79)
Chart 3	4.00 (1.52)	5.24 (1.58)	4.30 (1.68)	3.31 (1.71)
Chart 4	4.16 (1.67)	5.27 (1.62)	4.39 (1.84)	3.17 (1.82)

Table 18: TLX analysis of study 5: Enhanced Strength Graph

Condition	TLX1	TLX4	TLX5	TLX6
	M (SD)	M (SD)	M (SD)	M (SD)
Graph 1	4.34 (1.62)	5.13 (1.53)	4.38 (1.77)	3.24 (1.71)
Graph 2	4.45 (1.47)	5.18 (1.43)	4.73 (1.53)	3.28 (1.72)
Graph 3	4.17 (1.64)	4.98 (1.59)	4.21 (1.70)	3.36 (1.67)
Graph 4	4.20 (1.64)	5.01 (1.35)	4.53 (1.54)	3.28 (1.70)

8.4 SUMMARY AND DISCUSSION

In study 5, I ran a crowd-sourced online user study (study 5) through Amazon Mechanical Turk. A total of 400 and 417 participants were recruited in the study. The study helps to identify the effective design of *Two Way Bar Chart* and *Strength Graph*. I conducted a 2 by 2 design of the interface. In the group of *Two Way Bar Chart*, I controlled the conditions of the number of terms and the order methods. In the group of *Strength Graph*, I controlled the conditions of edge thickness and node size. The experiment results indicated the effectiveness of the *Chart 2* (60 terms with the order by individual relevance) and Graph 2 (enabled edge thickness and disabled node size) design. The two improved interfaces will be adopted in the later experiments as the explanation interfaces for *publication similarity* and *co-authorship similarity*.

9.0 CONTROLLABILITY AND EXPLAINABILITY IN A HYBRID SOCIAL RECOMMENDER SYSTEM

In this chapter, I report my exploration on bringing evaluated controllable user interface (pre-study) and explainable user interface (study 1 to 3) to a hybrid social recommender system. The combined interface has been integrated into the study 2 for testing the interaction between explanation visualizations in an interactive recommender system. A common limitation for the pre-study, study 1 & 2 is the small scale of human subjects, which limited the statistical power of the analyzed result. In this section, I intend to provide a large scale, lab-controlled human subject experiment (N=50) that to confirm the affirm the finding on user interface design and user performance, i.e., study 6.

9.1 INTRODUCTION

To explore the value of controllability and explainability in recommender context, I developed a social recommender interface for attendees of academic conferences. The system called *Relevance Tuner+* was implemented as a component of a conference support system - Conference Navigator [10]. It is an extension of my early controllable recommender system RelevanceTuner with several explanation-focused features [132, 136]. The original RelevanceTuner allowed users to "tune" (re-rank) the list of recommended co-attendees to meet using four sliders (see Fig 35, Section A). Each slider controlled the importance of one of the four source recommender engines of the hybrid recommender engine - *publication similarity*, *topic similarity*, *co-authorship similarity*, and *interest similarity*. By moving the slider to the right, the relative importance of the component could be increased; by moving

it to the left, the importance decreases. The final impact of a specific source on the ranking of a specific co-attendee was determined by both the relevance of this attendee along this source (for example, the similarity of her publication to the publications of the target user) and the current importance of the source selected by the user (e.g., importance of the Publication Similarity source). To make this hybrid fusion process more transparent, the RelevanceTuner interface used colored stackable bar visualization to show how the total relevance score of a recommended attendees was composed of source-level relevance scores (see Fig 35, Section B). More information about recommended people could be obtained using links in Section D. The RelevanceTuner+ further expanded this controllable interface with a *explanation icon* shown next to each recommendation in Section C, which could be used to *inspect* the relevance of this recommendation. Following a click on the explanation icon, a window will pop up (see Fig 36) to show a clickable explanation table, presenting visual explanations of calculated relevance for the selected recommendation engine.

The pop-up explanation window provided access to four explanation interfaces through color-coordinated buttons. Each interface focused on explaining the recommendation score delivered by the corresponding component of the hybrid engine: a) Two-way Bar Chart (for *Publication similarity*) showed the mutual relationship between two scholars publication terms and term frequency.; b) Topical Radar (for *Topic Similarity*) presented the top topics of the user and compared them with the selected attendee; c) Strength Graph (for *Co-Authorship similarity*) explained the strength of social connections by presenting the co-authorship network connecting the recommender attendee to the target user (nodes and edges are representing authors and co-authorship, respectively); d) Venn Tags (for *CN3 Interest similarity*) visualized the similarity of co-bookmarked, or co-followed item in the form of a Venn diagram. The design can be found in Fig 37. The specific visualization was constructed through a participatory design process presented in my previous study [134].

Relevance Tuner+ enhanced recommender system transparency in three ways. First, the user can inspect the static colored stackable bar to understand the fused relevance score in a multi-relevance context. Second, the user can influence the fusion by changing the weighs and observing the immediate changes of the colored relevance bars and item ranking. This controllability makes the fusion process transparent. Third, the user can access more

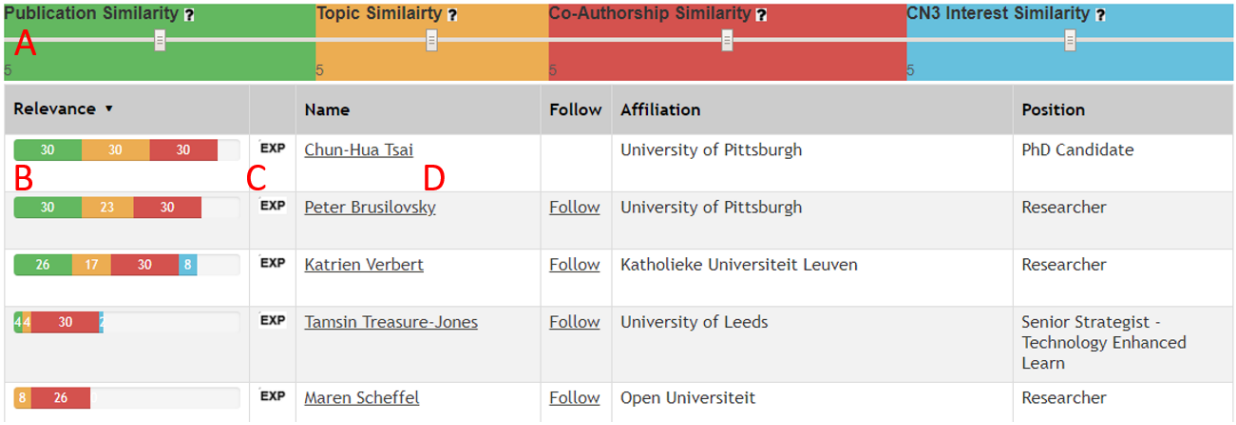


Figure 35: The design of Relevance Tuner+: (A) relevance sliders; (B) stackable score bar; (C) explanation icon; (D) user profiles.

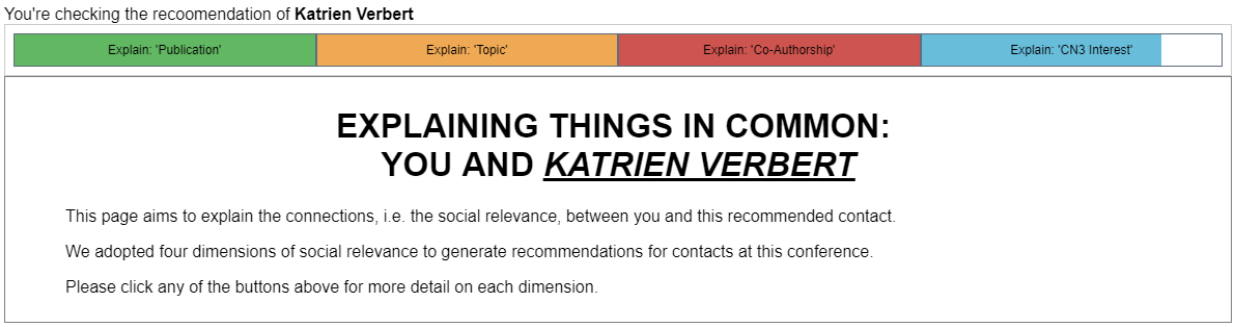


Figure 36: The pop-up window of clicking the *explanation icon* (shown in Figure 35 Section C).

information about the reasoning process by clicking the explanation icons. In the design of *Relevance Tuner+*, I attempted to separate the fusion process transparency from the explanation of each relevance obtained by clicking the explanation icon. To be more specific, I intend to gain system transparency in the aspect of the recommendation process, as well as the reasoning process.

To ensure the soundness of RelevanceTuner+, all its components were separately evaluated in a sequence of studies. The evaluation of the controllable slider was presented in my

previous study [132]. The findings showed that the visual interfaces significantly reduced the information search efforts tasks and helped users to perceive recommendation quality as well as an improvement in overall user satisfaction. The design of the explanation icon was discussed in my previous studies [136], the finding indicated the effectiveness of the proposed explanation models and a significant improvement in the perception of explainability, but I found providing controllability, and explainability was complementary that the users may not adopt both of the functions together. The visual interfaces for the four source explanation models were selected through a stage-based participatory design process to discuss the user preference of the interface prototypes [134] interactively. I further improved the design by a performance-focused evaluation. The result suggests that the user-preferred interface may not guarantee the same level of performance. To resolve the disagreements between designs favored by user preferences and efficiency evaluation, I conducted a user study [135] to determine the best designs for the recommendation models where the choice of the top design was not evident after the two earlier studies. The following section reviews the final version of the source recommendation model visualizations in more detail.

9.1.1 Explaining Recommendation Models

A total of four recommender models were introduced in RelevanceTuner+: *1) Publication Similarity:* cosine similarity of users' publication text; *2) Topic Similarity:* topic modeling similarity of research interests (topics); *3) Co-Authorship Similarity:* the degree of network distance, based on a shared co-authorship network; *4) CN3 Interest Similarity:* the number of papers co-bookmarked, as well as the authors co-followed. The detailed design of the explanations can be found here [134].

9.1.1.1 Publication Similarity: This similarity was determined by the degree of text similarity between two scholars' publication vectors using cosine similarity. I applied tf-idf to create the vector with a word frequency upper bound of 0.5 and a lower bound of 0.01 to eliminate both common and rarely used words. In this model, the key components were the *terms* of the paper title and abstract as well as its *term frequency*.

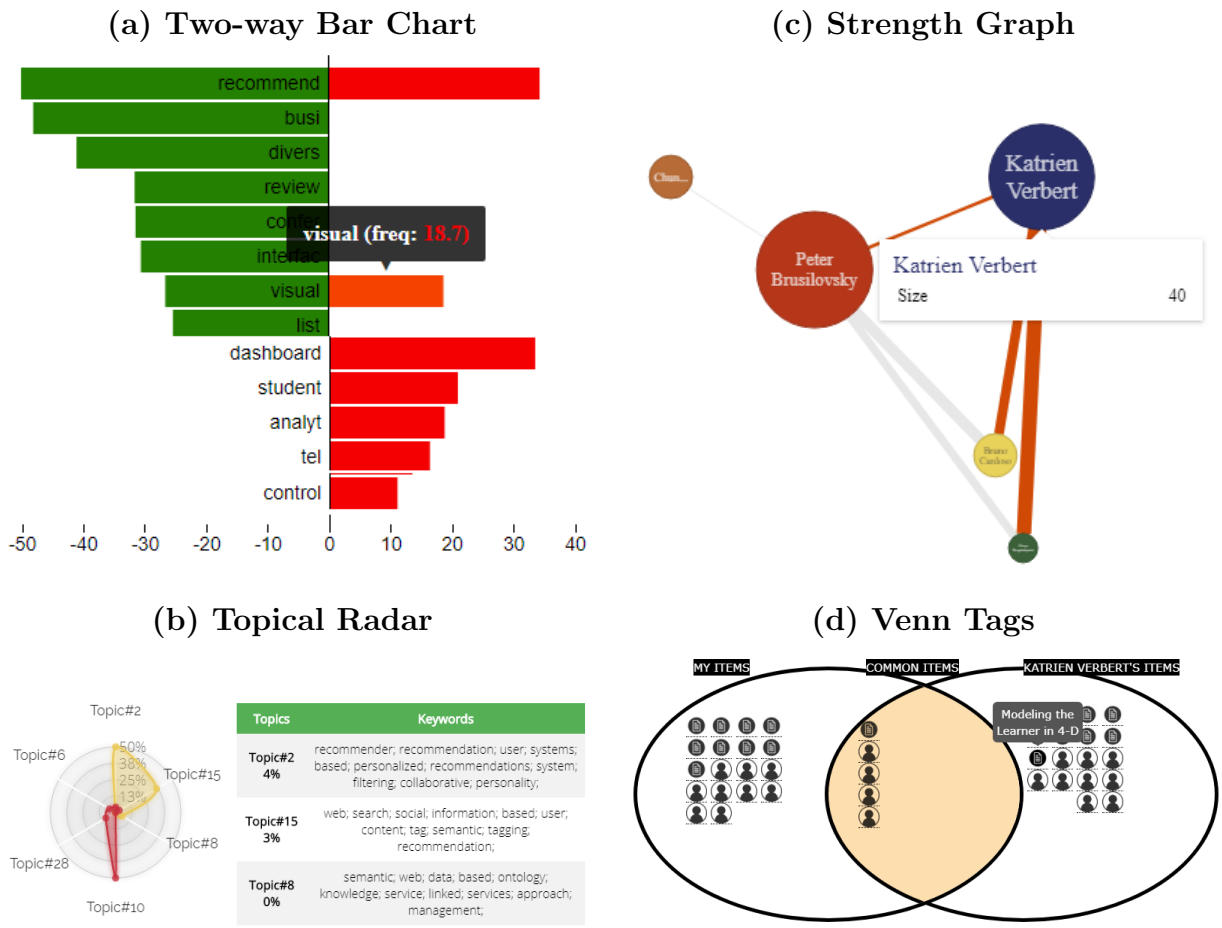


Figure 37: The explanation interfaces of study 6: (a) Two-Way bar chart for publication similarity; (b) Topical Radar for topic similarity; (c) Enhanced strength network for co-authorship network; (d) Venn Tag for CN3 interest similarity.

On the basis of my studies [134, 135] I adopted a *Two-Way Bar Chart* visualization as an approach to explaining the text-level similarity between the publication of the user and the attendee (Fig 37a). The visualization presented the mutual relationship of two scholars' publication terms and term frequency, i.e., one scholar in positive and the other scholar on a negative scale. This visualization presented the *terms* of the paper title and abstract. The bar length indicates the term frequency in the documents. The user's terms and the attendee's terms are presented on the left and right, respectively. The user can inspect the

words-in-common through the term appear in both sides, e.g., the term *visual* in Fig 37a means the term appeared the both of the scholars' publications. I ran a crowdsourcing study to determine the setting of the bar chat. Based on the study result, I choose 30 terms (versus 60 terms) ordered by individual relevance (versus the sum of relevance).

9.1.1.2 Topic Similarity: This similarity was determined by matching research interests using topic modeling. I used latent Dirichlet allocation (LDA) to attribute collected terms from publications to one of the topics. I chose 30 topics to build the topic model for all scholars. Based on the model, I then calculated the topic similarity between any two scholars. The key components were the *research topics* and the *topical words* of each research topic [148].

I presented research topics in a radar chart and the topical words of each research topic in table [148]. The visualization design can be found in Fig 37b. The radar chart was presented on the left side. I selected the top 5 (ranked by *beta* value from a total of 30 topics) topics of the user and compared them with the other scholar. A table with topical words was presented in the right so that the user can inspect the context of each research topic. I found this design is effective based on the user study of [136]. Based on the study result, the users were able to achieve 97% correct rate of sorting multiple recommendation models, solely using the visualization.

9.1.1.3 Co-Authorship Similarity: This similarity approximated the co-authorship network distance between the source and recommended users. For each pair of the scholar, I tried to find six possible paths for connecting them, based on their co-authorship relationships. The network distance is determined by the average distance of the six paths. The key components were the *coauthors* (as nodes), *coauthorship* (as edges) and the *distance of connection the two scholars*.

I presented a co-authorship network in a path graph [131]. The visualization design can be found in Fig 37c. For connecting the user (yellow circle on the left) to the attendee (red color in the right), I tried to find six possible paths (one shortest and five alternatives) by direct and in-direct co-authorship. In my original design [135], I found the user were failed

to use this visualization in sorting the recommendations. I then ran a crowdsourcing study to refine the network design. The study was a 2x2 factor design that has four conditions: edge thickness as the relevance between the connected nodes (i.e., the co-authored papers between two scholars) and node size as the number of papers (i.e., to make the node size presented the additional number of paper information). Based on the study result, I decided to set 1) node size as the number of papers; 2) edge thickness as the number of co-authored papers. The improved design has been shown effective in getting a 60% correct rate of sorting recommendations by relevance.

9.1.1.4 Interest Similarity: This similarity was determined by the number of co-bookmarked conference papers and co-connected authors in the conference support social system Conference Navigator (CN3). I used the number of shared items as the CN3 interest similarity. The key component is the shared *conference papers* and authors.

I presented co-bookmarked papers in a design of *Venn Tags* (shown in Fig 37d). The study of [75, 104] has pointed out the user preferred the Venn diagram as an explanation in a recommender system. The interface shown in Fig 37d: Venn Tags, I implemented the same idea with the bookmarked items. The idea is to present the bookmarked item, using an icon, in the Venn diagram. The two sides are the bookmarked item belong to one party. The co-bookmarked or co-followed item will be placed in the middle. The users can hover the icon for detail information, i.e., paper title or author name. I found this design is effective based on the user study of [136]. Based on the study result, the users were able to achieve 93% correct rate of sorting multiple recommendation models solely using the visualization.

9.2 SETUP (STUDY 6)

In study 6, I attempted to explore the effects of controllability and explainability in a realistic recommendation context. This goal was facilitated by the practical nature of RelevanceExplorer+, which was implemented and used as a social recommendation component or a popular conference support system Conference Navigator 3 (CN3), which has been used at

Table 19: Meta-data of the UMAP dataset

	UMAP 2015	UMAP 2016	UMAP 2017	UMAP 2018
Number of Papers	143	129	168	108
Number of Authors	231	305	345	289
Number of Attendees	116	115	151	131
Number of Bookmarks	664	660	714	342
Assigned Interface	BASE	CONT	EXPL	FULL

many research conferences [10]. However, the real conference context makes it impossible to run a reliable, controlled study. While in my past work, I did explore earlier versions of RelevanceExplorer in conference-based field studies [132], the needs of a randomized controlled study caused us to perform this study outside of a real conference context. However, I used data from real conferences and engaged experienced graduate students who were a close approximation to real conference participants. The study used different scenario-based information tasks and data-driven analysis methods.

9.2.1 Data and Participants

The recommendations produced by all four recommendation models are based on data collected by the CN3 system. I used real data from UMAP Conferences, from the year 2015 to 2018. UMAP is the premier international conference for researchers working on *Personalized Recommender Systems*, *Technology-Enhanced Adaptive Learning*, *Personalized Social Web Adaptive Hypermedia* and *the Semantic Web*. The dataset contained conference proceeding data, including conference papers (author, title and abstract), author and attendee list (name, published papers, affiliation, and position) and user feedback (bookmarks on papers and authors). As in the real conference context, my study used recommendation models to provide a social recommendation, i.e., recommending conference attendees to meet. The summary description of the UMAP dataset can be found in Table 19. Subjects were

recruited by emails and ads posted at the building of the School of Computing and Information, University of Pittsburgh. The promotional emails were sent to mailing lists of all doctoral students and summer registered master students. The main requirement was that the subjects should be able to perform an information search on web application as well as were majored in Library Science (LIS), Information Sciences (IS), Telecommunications (MST), Computer Sciences (CS) and the Intelligent Systems Program (ISP) at the School of Computing and Information. Each participant received USD$20 compensation and signed an informed consent form.

I conduct a power analysis to determine the size of the participants. A calculation of power analysis involves the following four parameters. 1) Alpha (α): it is a cut-off p-value that indicates the threshold probability for rejecting the null hypothesis (Type I error rate), the $\alpha = 0.05$. 2) Power: it is the probability of finding a true effect - the probability of failing to reject the null hypothesis under the alternative hypothesis (Type II error rate). I choose the power value as 0.8. 3) Effect size: it is the expected effect size, which refers to the expected correlation coefficient in this case. The goal is to find medium-sized effects in which the value is as 0.35. 4) Sample size N: this value determines the 'participant's size to maintain statistical power. I used *G*Power* software to calculate the sample size, which results in the calculation result that a total sample size of 49. The sample size supports the actual statistical power of 80%.

Based on the power analysis, a total of 50 participants (N=50) were recruited for the user study, from May 30 to June 16, 2019. The subjects were 28 males and 22 females whose ages ranged from 22 to 44 (M=28.82, SD=4.83). A total of 22 Master's students joined the study, included 21 IS and 1 MST majors. There were 28 doctoral students, included 18 IS, 3 LIS, 2 CS, and 5 ISP majors. All doctoral students had at least one publication and one-time conference attending experience, but no Master's students had any publication or ever attended any conference. Subjects took between 52 and 192 minutes (M=106.05, SD=28.80) to complete the study.

9.2.2 Experiment Design and Procedure

In this study, I explored *RelevanceTuner+*, a controllable and explainable user interface for the social recommendation. To access the separate and combined value of its controllability and explainability features, I used a 2x2 experimental design with four conditions. There were two intervention in this design: enable/disable relevance sliders (Section A in Fig 35) and enable/disable explanation icon (Section C in Fig 35). The four conditions produced by these interventions are: Baseline interface (**BASE**) with both slider and explanation icon disabled; Controllable interface (**CONT**) with slider enabled; Explainable interface (**EXPL**) with explanation icon enabled; Full interface (**FULL**) with both slider and explanation icon enabled. Each condition was used with a specific year of UMAP conference data (see Table 19). I followed the within-subject design, so all participants were asked to use each interface for one training and two study tasks and to fill out a post-stage questionnaire at the end of their work with each interface. At the end of the study, participants were asked to fill a post-study questionnaire. To minimize the learning effect and bias, I followed a Latin square design to balance the conditions that appeared to each participant. The study procedure can be summarized as followed steps:

1. Pre-study questionnaire (for user preference elicitation)
2. Interface 1: training session, task 1, task 2
3. Post-stage questionnaire
4. Interface 2: training session, task 1, task 2
5. Post-stage questionnaire
6. Interface 3: training session, task 1, task 2
7. Post-stage questionnaire
8. Interface 4: training session, task 1, task 2
9. Post-stage questionnaire
10. Post-study questionnaire

User Preference Elicitation: In the real-world conference scenario, most of the event attendees have some publications that the system can use to generate social recommendations. However, in my study, only a part of the subjects have ever published papers. To

better control the data sparsity, I asked the participants to complete a pre-study questionnaire before the user study. The pre-study questionnaire was aimed to conduct the user preference elicitation process so I can generate personalized social recommendations. The pre-study questionnaire has four questions. First, I asked subjects to pick five preferred research topics (out of 30 topics) that generated by the *Topic Similarity*. Each topic came with hundreds of *topical keywords*. I listed the top 10 keywords to let the user understand the context of each topic. For example, "*Topic#1: behavior, user, cultural, networks, social, world, web, differences, mixed, models, effect*". I used the selected topics for matching *Topic Similarity*. The topical keywords of the chosen topic were then used to calculate *Publication Similarity*. Second, I asked subjects to select five *mentors* from the faculty directory of the School of Computing and Information. I filtered 18 professors (out of 41 tenure-stream professors) whose research interest close to the topics of the UMAP conference. The selected mentors will be treated as the subjects' *coauthors* for calculating the co-authorship similarity. Third, I asked subjects to select three favorite papers from each year's UMAP top 10 bookmarked papers (a total of 12 papers). The data was used to calculate *CN3 Interest Similarity*.

In the study, the subjects were told to act as a researcher who is attending the conference. They were requested to pick a suitable candidate among conference attendees, based on their best judgment. Participants were given one training session and two information search tasks for each interface. In the training session, I urged the user study participants to follow a few steps so they have a chance to familiarize the system. The instrumentation details are listed below.

Training Session:

1. Inspect the definition of "publication similarity" (click on the question mark next to it) and read the text that appears.
2. Re-tune one or more sliders and inspect how adjusting that changes the stackable color bar.
3. Click two scholar names and check their publications.
4. Click two "Explanation Icon" and inspect the four explanation functions in them.

5. Follow any four scholars by clicking the Follow button, you will need to provide a reason for your selection. Please type "test" as your answer.

To be noted, in step 1, the inspecting similarity were rotated across conditions to cover all four similarities. Step 2 is only appeared in *CONT* and *FULL* interface. Step 4 is only appeared in *EXPL* and *FULL* interface. In step 5, a pop-up box will present to collect user feedback/comments when the user clicked on "follow" button (see Fig 35, "Follow" column).

The study tasks 1 and 2 were explicitly designed as a relevance-oriented and diverse-oriented information search tasks, respectively. I tried to make the assigned tasks as realistic that could be naturally pursued by attendees at research conferences. For relevance-oriented task (Task 1), I expect to see the user to coordinate multiple relevance aspects and find out the desired candidates efficiently. In contract, the diverse-oriented task (Task 2), I expect the user to select candidates from different relevance aspects, which diverse their selection. Both of the tasks asked the user to "follow" four scholars from the conference attendee list, based on different criteria. To promote the function usage and advance the information need, in each "following", I asked the user to "justify" their choice by typing a couple of sentences about the reasons they decided to follow the chosen scholar (same as the Step 5 above).

Task 1: Find Advisor/Mentor

- If you plan to pursue a doctoral degree after your current degree program, it is an excellent opportunity to find your prospective advisor or mentor at the conference. For this task, you will select scholars to follow as potential advisors/mentors.
- Please "follow" four scholars whose work in more relevant to your research interest(s). The ideal candidates will be scholars who the system identified as more connected to your chosen SCI professors (so they can provide you a strong recommendation letter).
- You are also expected to justify your selections (for example, to the Ph.D. admission committee), so it is important to pay attention to why do you make the selection.

Task 2: Find a Guest Speaker

- You are asked to invite guest speakers for an academic seminar at your home school. The main seminar theme is to encourage the inclusion of different types of research and community.

- Please Please "follow" four scholars from the conference whose work matches the seminar theme. The ideal candidates are: 1) those who can represent the inclusion of different types of research, i.e., those scholars are expected to work on different research topics; 2) those who are "less" connected to your chosen SCI professors.
- You are also expected to justify your selections (for example, to the dean's office), so it is important to pay attention to why do you make the selection.

9.2.3 Measurements

I used both subjective metrics and objective metrics to measure the effectiveness of the controllable and explainable interfaces. The subjective measures were captured by questionnaires. I used the existing constructs (groups of questions) from the works of [108, 70, 71]. In this study, I collected the user feedback in seven constructs: Perceived Recommendation Quality/Accuracy(Q), Perceived Recommendation Diversity/Variety(D), Perceived Control (C), Perceived System Effectiveness (E), Perceived Trust (T), Perceived Transparency (P) and Satisfaction (S). All these constructs were collected in the stage of the post-stage survey, i.e., after the participants interacted with each experiment condition. The questions are listed below, and the survey results were shown in Table 22.

- Construct Q: Perceived Recommendation Quality/Accuracy
  - Q1: The recommender was providing good recommendations.
  - Q2: I liked the recommendations provided by the system.
  - Q3: The recommended scholars fitted my preference.
  - Q4: I did not like any of the recommended scholars.
- Construct D: Perceived Recommendation Diversity/Variety
  - D1: The recommender helped me discover new contacts at conference.
  - D2: The scholars that recommended to me are diverse.
  - D3: The list of recommendations included scholars of many different research areas.
  - D4: The list of recommendations was very similar.
- Construct C: Perceived Control
  - C1: The recommender allows me to modify my preference.

Table 20: The results of post-study survey of study 6: I collected three constructs of personal characteristics (PC) and two constructs of system-specific characteristics (SC).

Construct	Factor		Score M (SD)
General Trust in Technology (PC1)	1	I feel technology never works.	1.34 (0.65)
	2	I'm less confident when I use technology.	1.76 (0.93)
	3	The usefulness of technology is highly overrated.	2.82 (1.49)
	4	Technology may cause harm to people.	3.52 (1.95)
		Average	2.36 (1.59)
User Characteristics: Scholarly Expertise (PC2)	1	Compared to my peers, I have a lot of collaborators or research experiences.	3.84 (1.62)
	2	Compared to my peers, I am an expert on the subject of the conference.	3.40 (1.51)
	3	I only know a few scholars at the conference.	5.04 (1.51)
	4	I frequently attend academic conferences.	3.50 (1.72)
		Average	3.94 (1,71)
General Acceptance of Diversity (PC3)	1	I'd like to see scholars with dissimilar research interests.	6.04 (1.02)
	2	I would be satisfied if I was recommended unfamiliar items.	4.54 (1.63)
	3	I think the recommendation should cover the scholars from different research areas.	5.96 (1.41)
	4	I like to see the recommendations beyond my interests.	5.48 (1.32)
		Average	5.05 (1.48)
System-specific Privacy Concern (SC1)	1	I'm afraid the system discloses private information about me.	3.26 (1.84)
	2	Personalized recommender usually invades my privacy.	5.56 (1.83)
	3	I'm uncomfortable providing private data even if it helps me to receive better recommendations	3.30 (1.75)
	4	I think the recommender system should respect the confidentiality of my data.	6.00 (1.37)
		Average	4.03 (2.04)
System-specific Familiarity and Understanding (SC2)	1	Compared to my peers, I am familiar with the technology of recommender systems.	4.80 (1.47)
	2	I feel comfortable to solve some mathematical or equation questions.	5.74 (1.32)
	3	I am confident when I first time interacts with a new information system.	5.26 (1.33)
	4	I am familiar with programming language.	6.02 (1.31)
		Average	5.45 (1.43)

  - C2: I became familiar with the system very quickly.
  - C3: The layout of the recommender interface is adequate.
  - C4: The recommender helped me to make the following decision faster.
- Construct E: Perceived System Effectiveness
  - E1: Using the system is a pleasant experience.
  - E2: I make better choices with the recommender.
  - E3: I can find better items using the recommender.
  - E4: I feel bored when Im using the recommender.
- Construct T: Perceived Trust
  - T1: I am convinced by the scholar recommended to me.
  - T2: I am confident I will like the items recommended to me.
  - T3: The recommender made me more confident about my selection/decision.
  - T4: The recommender can be trusted.
- Construct P: Perceived Transparency
  - P1: The provided information was sufficient for me to make a good decision.
  - P2: The recommender explained why the scholars were recommended to me.
  - P3: I understood why the scholars were recommended to me.
- Construct S: Satisfaction
  - S1: I will use this recommender again.
  - S2: I will tell my friends about this recommender.
  - S3: Overall, I am satisfied with the recommender.
  - S4: The recommender helped me find the ideal contacts at the conference.

I further collect three constructs of personal characteristics (PC) and two constructs of system-specific characteristics (SC) in the post-study survey. These constructs included General Trust in Technology (PC1), Scholarly Expertise (PC2), General Acceptance of Diversity (PC3), System-specific Privacy Concern (SC1), System-specific Familiarity and Understanding (SC2). The constructs help to understand the user personality and expertise that were positively correlated with user perceptions [115]. The detail questions and survey results of PC and SC constructs can be found in Table 20.

I used objective metrics to measure user engagement and recommendation quality. In all four study interfaces, the user can *inspect* the recommendation by clicking on different links and interface elements. I logged five user action in this user study, included click on *Learn more Scholar Name*, *Relevance Tuner*, *Explanation Icon* and *Explanation Tabs*. I also logged the *Time Spent* to measure the time used in each task. In addition to user engagement measures, I adopted four metrics to evaluate the recommendation quality [104]: *1) Top N:* to measure the accuracy of a list of top k recommendation. *2) Mean Reciprocal Rank (MRR):* MRR was determined by the ranking position of the first relevant element is matched [**?**]. *3) Normalized Discounted Cumulative Gain (NDCG):* This metric penalized the lower-ranked relevant items, which approached the relevant item to be ranked in the top positions [88]. 4) alpha-NDCG: it is relevant to NDCG but as used to measure for diversified search, where it is appreciated by the number of covered intents [22]. All these objective metrics provided a sketch and description of the user interaction with the proposed interfaces.

9.3 RESULTS

9.3.1 Action Analysis

Table 21 presents system usage (number of clicks and time spent) of two study tasks in four different conditions. Most importantly, the data indicates that the participants extensively used both control tuners and an explanation icon when these options were provided. There were 42 (out of 50) users who used the tuners to solve the tasks and 39 (out of 50) users who used the explanation when solving the tasks. The users usually inspected 2-3 explanation tabs each time after clicking the explanation icon. I found that the users used the tuners and explanation icon more intensively in the first task than the second task. It is consistent with my previous finding [132] that the user engagement rate is higher when the user interacts with an intelligent interface for the first time. The time spent supports the argument that it took less time to solve the second task. The analysis indicates that the need to examine details about a specific candidate user by clicking *Scholar Name* is noticeably decreased

Table 21: User action summary of study 6: the table shows the user interaction statistics while performing each of the two tasks using four interfaces. (UC: User Count)

	Actions	BASE (S1)		CONT (S2)		EXPL (S3)		FULL (S4)	
		M(SE)	UC	M(SE)	UC	M(SE)	UC	M(SE)	UC
TASK 1	Learn More	0.72 (1.52)	15	0.74 (1.58)	14	0.52 (0.97)	16	0.36 (0.98)	9
	Scholar Name	13.66 (10.80)	44	15.12 (16.03)	43	10.52 (14.24)	41	11.72 (17.10)	40
	Tuners	-	-	30.26 (39.97)	42	-	-	22.38 (25.67)	42
	Exp. Icon	-	-	-	-	7.28 (8.66)	39	7.42 (10.74)	35
	Exp. Tabs	-	-	-	-	18.56 (23.63)	37	15.22 (18.44)	34
	Time Spent (in Seconds)	517.24 (343.80)	50	633.46 (682.92)	50	613.84 (409.65)	50	533.32 (303.71)	50
TASK 2	Learn More	0.24 (0.82)	6	0.22 (0.54)	8	0.26 (0.87)	6	0.50 (1.24)	9
	Scholar Name	14.36 (14.35)	44	14.42 (11.89)	45	9.92 (9.86)	43	14.26 (16.79)	41
	Tuners	-	-	19.34 (21.84)	44	-	-	20.22 (25.81)	40
	Exp. Icon	-	-	-	-	5.56 (7.70)	29	5.64 (8.26)	32
	Exp. Tabs	-	-	-	-	11.26 (12.88)	29	12.88 (21.98)	31
	Time Spent (in Seconds)	418.28 (310.92)	50	479.24 (341.98)	50	460.12 (289.90)	50	481.56 (299.24)	50

when explanations are available. i.e., users clicks *Scholar Name* least frequently in the *EXPL* interface and most frequently in the *CONT* interface. It hints that the explanations do help to understand the match. However, as I tested the user interaction parameters between the four interfaces using a two-way ANOVA, but I did not find significance for *Learn More*, *Scholar Name* and *Time Spent* metrics.

I further analyzed the user interaction of the control and explanation functions. Fig 38 presents the average re-weighting value of *CONT* and *FULL* interfaces through four box-plots. The average re-weighting value was calculated by $\frac{\sum_{i=1}^{n} Slider_s(n)}{n}$, where n represents number of click on the slider s and $S_s(n)$ represents the selected weight for each slider adjustment attempt (from 0 to 10). The usage analysis shows that the users tend to increase the relevance weighs from its default weighting value of 5 in relevance-oriented Task 1 when no compromise between sources is necessary and decrease the weights in diversity-oriented Task 2 when a balance between sources has to be found. I can observe a similar pattern across all four relevance sliders and both of the interfaces. In other words, the users tend to *enhance* the source importance while seeking similar items, but to *reduce* the importance while seeking dissimilar items. The effect was most visible in adjusting publication and topic similarity, which were primary targets in both tasks and least pronounced for co-authorship similarity in Task 1, where it was a secondary (although positive) consideration. However, I found the effect is less obvious in the group of co-authorship similarity. This data hints that the users understood the purpose of the sliders and used them in accordance with the nature of the assigned information tasks.

9.3.2 Recommendation Quality

To measure recommendation quality, I treated user selections as ground truth, assuming that the graduate-level users are the best judges on who is the right person for him or her to meet. In essence, all metrics used in my evaluation attempted to measure in different ways to what extent a specific version of the system, with or without the help of the user-driven tuning, is able to place relevant items in the higher ranks of the recommended people list. Given the well-known tendency of search and recommender system users to focus on the top

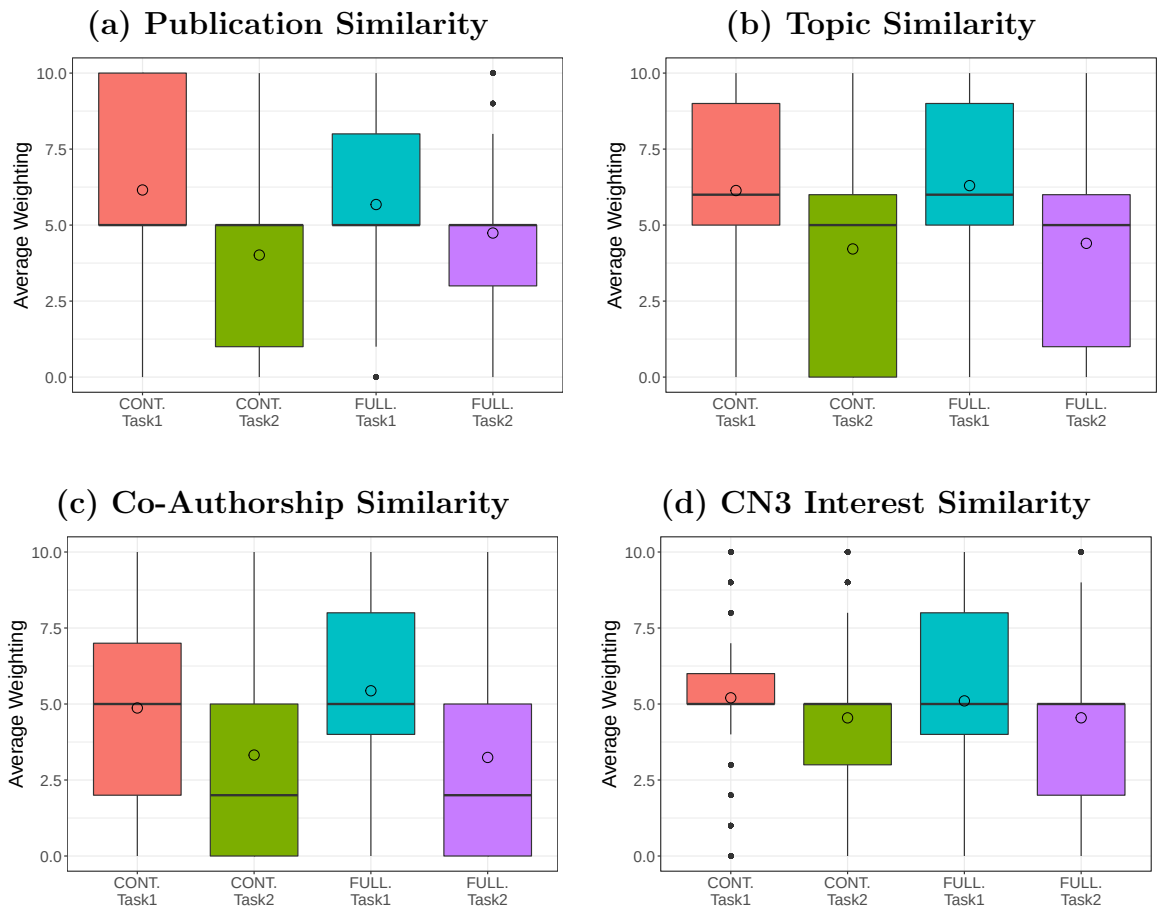


Figure 38: The average re-weighting value of *CONT* and *FULL* interfaces.

positions in the ranked list, a system that helps the user to find relevant people high the ranked list offers a better recommendation quality than the one that pushes the user to go to the lower ranks.

I present the analysis of objective measures in Fig. 39 showing curves for each measure taken at points $k = 10$ to $k = 100$. A one-way between-subjects ANOVA was conducted to compare the effect of user interfaces on recommendation quality in *BASE*, *CONT*, *EXPL* and *FULL* conditions. In task 1, I did not find any significant differences between the four conditions. That is, in relevance-oriented Task 1, I did not observe that between the four interfaces, the users selected recommendations from significantly different ranking positions

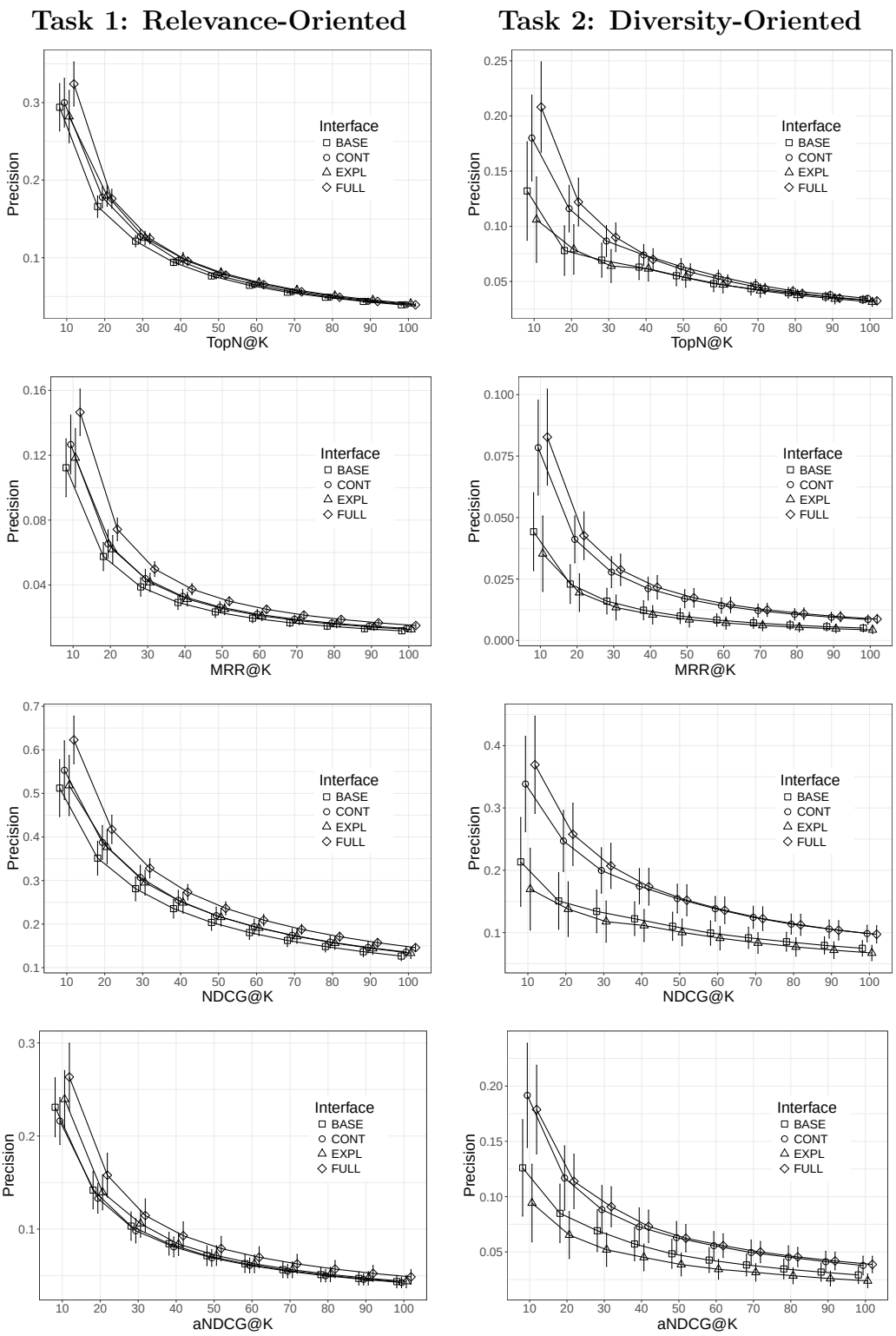


Figure 39: The recommendation quality analysis of study 6: Top-N, MRR, NDCG and alpha-NDCG.

as measured by TopN, MRR, and NDCG metrics. However, I can still see the *FULL* tends to outperform all the other three conditions as measured by all performance measures. The result hints that are providing controllable and explainable interfaces can increase the recommendation quality, i.e., the useful recommendations selected by the users tend to be elevated to the higher ranks. In task 1, it is also reasonable to see there is no difference in alpha-NDCG metric since I didn't expect the users to select a set of diverse recommendations.

In task 2, there was a significant effect of interface on recommendation quality metrics at the $p < 0.05$ level for the four conditions. I found significance on TopN@10 [F(3, 196) = 5.12, p = 0.001], MRR@10 [F(3, 196) = 7.44, $p < 0.001$] and NDCG@10 [F(3, 196) = 7.00, $p < 0.001$] metrics. I then conducted a post-test due to the significance. Post hoc comparisons using the Tukey HSD test indicated that 1) the TopN mean score for the *FULL* condition (M = 0.20, SD = 0.14) was significantly different than the *BASE* (M = 0.13, SD = 0.15) and *EXPL* (M = 0.10, SD = 0.35) conditions. Moreover, the *CONT* condition (M = 0.18, SD = 0.13) did significantly differ from the *EXPL* (M = 0.10, SD = 0.35) condition; 2) the MRR mean score for the *FULL* condition (M = 0.08, SD = 0.06) was significantly different than the *BASE* (M = 0.04, SD = 0.05), *CONT* (M = 0.07, SD = 0.06) and *EXPL* (M = 0.03, SD = 0.5) conditions. Moreover, the *CONT* condition (M = 0.07, SD = 0.06) did significantly differ from the *BASE* (M = 0.04, SD = 0.05) condition. 3); the NDCG mean score for the *FULL* condition (M = 0.36, SD = 0.27) was significantly different than the *BASE* (M = 0.21, SD = 0.25) and *EXPL* (M = 0.16, SD = 0.23) conditions. Moreover, the *CONT* condition (M = 0.33, SD = 0.26) did significantly differ from the *BASE* (M = 0.21, SD = 0.25) and *EXPL* (M = 0.16, SD = 0.3) conditions. I found significance on alpha-NDCG@10 [F(3, 196) = 4.80, p = 0.002] metric as well. The alpha-NDCG mean score for the *FULL* condition (M = 0.17, SD = 0.14) was significantly different than the *EXPL* (M = 0.09, SD = 0.12) conditions. Moreover, the *CONT* condition (M = 0.19, SD = 0.16) did significantly differ from the *EXPL* (M = 0.09, SD = 0.12) condition.

To summarize, I found that in diversity-oriented task 2, providing controllability can increase the recommendation quality. It means that using the sliders, users could better tune the ranking to their preferences bringing most relevant items to the higher positions in the ranked list. However, the findings also indicate that providing explainability may not

help to increase the recommendation quality as measured by my metrics. As I observed, the addition of an explanation function leads the user to go deeper into the ranking list when inspecting and selecting recommendations, which is bad from the prospects of these metrics. I argue, however, that the reduction of quality in the presence of explanations is likely the effect of using traditional ranking-focused recommendation metrics, which were not designed to assess the explanation aspects and might neither reflect the true performance of a system nor the ultimate user satisfaction. It is known that when decision-support information is scarce, ranking is one of the most important factors to assess item relevance [64], causing people to prefer higher-ranked items. However, the presence of decision-support information in the form of explanations could have made the users more confident to explore the lower ranks and judge people positioned there as relevant to meet. In this study, I attempted to minimize the inherent bias of higher ranking by assessing the value of each condition using a combination of objective and subjective metrics.

At the same time, it is interesting to observe that selecting the recommendations beyond the top-ranks is not a guarantee that the users will achieve a higher selection diversity. In the finding of alpha-NDCG, I found a higher alpha-NDCG score in *CONT* and *FULL* conditions. The *EXPL* demonstrated the lowest alpha-NDCG score, which implied the users failed to select recommendations that covered sufficiently different research topics.

9.3.3 User Feedback Analysis

9.3.3.1 Interface Differences Table 22 shows the analysis of post-stage survey focused on differences between the interfaces. I built 7 constructs based on 27 questions, included *Perceived Recommendation Quality (Q)*, *Perceived Recommendation Diversity (D)*, *Perceived Control (C)*, *Perceived System Effectiveness (E)*, *Perceived Trust (T)*, *Perceived Transparency (P)* and *Satisfaction (S)*. All these constructs reflected the user experience after using the four interface conditions. A one-way between subjects ANOVA was conducted to compare the effect of user interfaces on user subjective feedback in *BASE*, *CONT*, *EXPL* and *FULL* conditions.

I found significance on all constructs, included *Perceived Recommendation Quality (Q)*

[F(3, 196) = 4.00, $p < 0.001$], *Perceived Recommendation Diversity (D)* [F(3, 196) = 2.96, p = 0.03], *Perceived Control (C)* [F(3, 196) = 12.82, $p < 0.001$], *Perceived System Effectiveness (E)* [F(3, 196) = 4.62, $p < 0.001$], *Perceived Trust (T)* [F(3, 196) = 9.77, $p < 0.001$], *Perceived Transparency (P)* [F(3, 196) = 17.63, $p < 0.001$] and *Satisfaction (S)* [F(3, 196) = 6.17, $p < 0.001$]. I then conducted a post-test for all constructs that shown significance. The Post-hoc comparisons using the Tukey HSD test was reported below:

1. The score of *Perceived Recommendation Quality (Q)* for the *FULL* condition (M = 4.72 SD = 0.68) was significantly higher than the *BASE* (M = 4.19, SD = 0.91) condition. However, the *CONT* (M = 4.52, SD = 0.76) and *EXPL* (M = 4.50, SD = 0.71) conditions did not significantly differ from the *BASE* condition.
2. The score of *Perceived Recommendation Diversity (D)* for the *FULL* condition (M = 5.24 SD = 0.78) was significantly higher than the *BASE* (M = 4.19, SD = 0.91) condition. However, the *CONT* (M = 5.10, SD = 0.97) and *EXPL* (M = 5.14, SD = 0.78) conditions did not significantly differ from the *BASE* condition.
3. The score of *Perceived Control (C)* for the *FULL* (M = 5.94 SD = 0.83) and *CONT* (M = 5.66, SD = 0.90) conditions were significantly higher than the *BASE* (M = 4.84, SD = 1.09) condition. Moreover, it was significantly higher for the *FULL* condition than the *EXPL* (M = 5.18 SD = 1.00) condition.
4. The score of *Perceived System Effectiveness (E)* for the *FULL* (M = 5.04 SD = 0.72) and *CONT* (M = 4.88, SD = 0.69) conditions were significantly higher than the *BASE* (M = 4.49, SD = 0.87) condition. Moreover, it was significantly higher for the *FULL* than the *EXPL* (M = 5.18 SD = 1.00) condition. However, the *EXPL* (M = 4.83, SD = 0.72) condition did not significantly differ from the *BASE* condition.
5. The score of *Perceived Trust (T)* for the *FULL* (M = 5.63 SD = 1.13), *CONT* (M = 5.25, SD = 1.06) and EXPL (M = 5.32, SD = 1.00) conditions were significantly higher than the *BASE* (M = 4.84, SD = 1.09) conditions.
6. The score of *Perceived Transparency (P)* for the *FULL* (M = 5.74 SD = 1.12), *CONT* (M = 4.81, SD = 1.27) and EXPL (M = 5.66, SD = 1.06) conditions were significantly higher than the *BASE* (M = 4.16, SD = 1.53) conditions. Moreover, it was significantly higher for the *FULL* and *EXPL* conditions than the *CONT* condition.

7. The score of *Satisfaction (S)* for the *FULL* (M = 5.60 SD = 1.27) and *CONT* (M = 5.39, SD = 1.06) conditions was significantly higher than the *BASE* (M = 4.58, SD = 1.4353) conditions.

To summarize, I found that each of my two novel features, control, and explanations, taken alone, tend to affect user experience positively, as measured by a range of experience-focused constructs, reaching significance for several metrics. As a whole, controllability seems to provide a slightly stronger impact generating higher scores and reaching a significant difference with the baseline in more metrics (4 vs. 1). At the same time, the perceived transparency is affected significantly stronger by adding explanations. Yet, it is together that these features could provide the strongest impact producing highest experience scores and reaching a significant difference with the baseline in all experience constructs.

9.3.3.2 Cohort Differences It is normal to assume the participants' academic background would affect the perception of using the systems. To examine this aspect of the study, I contrasted user perceptions between two academic groups, i.e. PHD group and MASTER group. I found significant differences between the two groups for all constructs, including *Perceived Recommendation Quality (Q)* [$F(1, 192) = 7.56$, $p < 0.001$], *Perceived Recommendation Diversity (D)* [$F(1, 192) = 11.46$, $p < 0.001$], *Perceived Control (C)* [$F(1, 192) = 7.08$, $p < 0.001$], *Perceived System Effectiveness (E)* [$F(1, 192) = 4.23$, $p < 0.001$], *Perceived Trust (T)* [$F(1, 192) = 20.89$, $p < 0.001$], *Perceived Transparency (P)* [$F(1, 192) = 15.13$, $p < 0.001$] and *Satisfaction (S)* [$F(1, 192) = 15.13$, $p < 0.001$]. I then conducted a post-test for all constructs that shown significance. The Post-hoc comparisons using the Tukey HSD test was reported below:

1. The score of *Perceived Recommendation Quality (Q)* for the *MASTER* group (M = 4.69 SD = 0.76) was significantly higher than the *PHD* (M = 4.30, SD = 0.77) group.
2. The score of *Perceived Recommendation Diversity (D)* for the *MASTER* group (M = 5.32 SD = 0.81) was significantly higher than the *PHD* (M = 4.84, SD = 0.86) group.
3. The score of *Perceived Control (C)* for the *MASTER* group (M = 5.61 SD = 1.00) was significantly higher than the *PHD* (M = 5.23, SD = 1.06) group.

Table 22: User feedback analysis: I found the *FULL* condition is significantly outperformed the BASE condition. Statistical significance level: (*) $p < 0.05$. (**) $p < 0.01$)

Construct	Factor	Interfaces			
		BASE M(SD)	CONT M(SD)	EXPL M(SD)	FULL M(SD)
Perceived Recommendation Quality (Q)	Q1	4.86 (1.48)	5.42 (0.99)	5.32 (1.25)	5.72 (1.08)
	Q2	4.96 (1.59)	5.26 (1.27)	5.34 (1.08)	5.76 (1.02)
	Q3	4.48 (1.46)	5.22 (1.25)	5.10 (1.16)	5.48 (1.29)
	Q4	2.46 (1.45)	2.20 (1.41)	2.26 (1.33)	1.92 (1.19)
	Ave	4.19 (0.91)	4.52 (0.76)	4.50 (0.71)	**4.72 (0.68)****
Perceived Recommendation Diversity/Variety (D)	D1	5.04 (1.51)	5.48 (1.38)	5.72 (1.12)	6.06 (1.05)
	D2	5.18 (1.20)	5.30 (5.36)	5.32 (1.23)	5.36 (1.54)
	D3	4.86 (1.67)	5.22 (1.44)	5.30 (1.26)	5.46 (1.18)
	D4	3.96 (1.61)	4.42 (1.42)	4.22 (1.63)	4.08 (1.66)
	Ave	4.76 (0.83)	5.10 (0.97)	5.14 (0.82)	**5.24 (0.78)****
Perceived Control (C)	C1	4.02 (2.27)	6.02 (1.05)	4.52 (2.09)	6.20 (0.94)
	C2	6.04 (1.17)	6.16 (1.07)	5.82 (1.27)	6.18 (0.91)
	C3	4.70 (1.60)	5.02 (1.54)	5.12 (1.45)	5.60 (1.34)
	C4	4.62 (1.80)	5.46 (1.23)	5.26 (1.39)	5.78 (1.35)
	Ave	4.84 (1.09)	**5.66 (0.90)****	5.18 (1.00)	**5.94 (0.83)****
Perceived System Effectiveness (E)	E1	4.42 (1.79)	5.26 (1.48)	5.10 (1.31)	5.50 (1.40)
	E2	4.84 (1.53)	5.48 (1.24)	5.60 (1.01)	5.84 (1.21)
	E3	4.92 (1.44)	5.58 (1.24)	5.34 (1.45)	5.82 (1.00)
	E4	3.80 (1.86)	3.20 (1.69)	3.28 (1.69)	3.02 (1.77)
	Ave	4.49 (0.87)	**4.88 (0.69)***	4.83 (0.72)	**5.04 (0.72)****
Perceived Trust (T)	T1	4.34 (1.67)	5.22 (1.20)	5.18 (1.38)	5.60 (1.24)
	T2	4.40 (1.48)	5.30 (1.24)	5.20 (1.24)	5.52 (1.34)
	T3	4.34 (1.67)	5.02 (1.37)	5.46 (1.14)	5.64 (1.34)
	T4	4.64 (1.39)	5.48 (1.01)	5.44 (1.07)	5.76 (1.28)
	Ave	4.43 (1.40)	5.25 (1.06)	5.32 (1.00)	**5.63 (1.13)****
Perceived Transparency (P)	P1	4.08 (1.80)	4.90 (1.51)	5.16 (1.34)	5.58 (1.23)
	P2	3.98 (1.84)	4.44 (1.78)	6.00 (1.17)	5.94 (1.39)
	P3	4.42 (1.72)	5.10 (1.46)	5.82 (1.40)	5.70 (1.40)
	Ave	4.16 (1.53)	**4.81 (1.27)***	**5.66 (1.06)****	**5.74 (1.12)****
Satisfaction (S)	S1	4.46 (1.70)	5.42 (1.27)	5.08 (1.50)	5.74 (1.42)
	S2	4.64 (1.78)	5.38 (1.45)	5.12 (1.43)	5.74 (1.48)
	S3	4.44 (1.57)	5.32 (1.16)	5.20 (1.27)	5.54 (1.32)
	S4	4.80 (1.34)	5.46 (0.99)	5.12 (1.33)	5.40 (1.39)
	Ave	4.58 (1.43)	**5.39 (1.06)****	5.13 (1.22)	**5.60 (1.27)****

4. The score of *Perceived System Effectiveness (E)* for the *MASTER* group (M = 4.97 SD = 0.79) was significantly higher than the *PHD* (M = 4.67, SD = 0.74) group.
5. The score of *Perceived Trust (T)* for the *MASTER* group (M = 5.50 SD = 1.17) was significantly higher than the *PHD* (M = 4.86, SD = 1.21) group.
6. The score of *Perceived Transparency (P)* for the *MASTER* group (M = 5.39 SD = 1.32) was significantly higher than the *PHD* (M = 4.83, SD = 1.44) group.
7. The score of *Satisfaction (S)* for the *MASTER* group (M = 5.47 SD = 1.23) was significantly higher than the *PHD* (M = 4.93, SD = 1.31) group.
8. I did not find interaction effect between the interfaces and degree groups for any of the constructs

The result shows a clear pattern: the MASTER group has higher user experience scores than the Ph.D. group. The findings demonstrate that personal factors (such as level of education, domain experience, and familiarity with technology) could significantly affect user perception of the system. Given the nature of the computer and information science field, the difference between groups in their technical knowledge (i.e., interfaces, recommender systems, the Web) is likely to be much smaller than in their domain knowledge (i.e., research, academia, publication, co-authorship, advising). It hints that an interactive and transparent recommender system could be of more value and importance to the users with lower domain knowledge for whom making their own decision in a less familiar domain could be a considerable challenge.

9.4 STRUCTURAL EQUATION MODELING

To build a complete understanding of the impact provided by my novel interface features on user experience, I conducted a structural equation model (SEM) analysis advocated in [70]. SEM analysis helps to explain the relationship between unobserved constructs (latent variables) using observable variables. For example, I may not be able to find perfect measurements that represent the user experience of using an intelligent system. However, I can adopt several items (questions) to measure user experience. In this analysis, I build

constructs through subjective measurements; I used the post-stage survey comprised of 16 questions that covered seven different user experience (UX) dimensions. I also combined three constructs of personal characteristics (PC) and three constructs of system-specific characteristics (SC).

In order to make sure the items (questions) in each construct are meaningful, I started with confirmatory factor analysis (CFA) and examined the construct validity. I tested the construct through two parameters: *Convergent validity* that makes sure the items in the construct are related and *Discriminant validity* that makes sure the unrelated items are really unrelated. The convergent validity of constructs was maintained by examining the average variance extracted (AVE) of each construct. In my analysis, with one exception, the AVEs of all constructs were higher than the recommended value of 0.50, indicating adequate convergent validity [71]. I had to remove one item from several constructs due to low variance, but the remaining items shared at least 48% of their variance with their designated construct. In order to ensure discriminant validity, I ascertained that the square root of the AVE for each construct was higher than the highest correlations of the construct with other constructs. I planned three sets of latent constructs: 6 subjective system aspects (SSA), one user experience (EXP) (satisfaction), three constructs of personal characteristic (PC), and three constructs of system-specific characteristics (SC). All statistics summarized below supports good convergent validity (AVE) and internal consistency (Cronbach's α).

- *SSA: Perceived Recommendation Quality (Q)*: 3 items (Q1, Q2, Q3). I removed items Q4 due to low variance (communality: 0.26) with the designated construct. ($AVE = 0.75$, $\sqrt{AVE} = 0.86$, $\alpha = 0.87$, largest correlation = 0.85)
- *SSA: Perceived Recommendation Diversity (D)*: 3 items (D1, D2, D3). I removed item D4 due to low variance (communality: 0.006) with the designated construct. ($AVE = 0.60$, $\sqrt{AVE} = 0.77$, $\alpha = 0.67$, largest correlation = 0.75)
- *SSA: Perceived Control (C)*: 4 items (C1, C2, C3, C4). While it is a popular construct, in my study the convergent validity was not adequate for this construct ($AVE < 0.5$). I discarded this construct in my analysis. ($AVE = 0.37$, $\sqrt{AVE} = 0.60$, $\alpha = 0.61$, largest correlation = 0.77)
- *SSA: Perceived System Effectiveness (E)*: 3 items (E1, E2, E3). I removed item E4

due to low variance (communality: 0.31) with the designated construct. ($AVE = 0.75$, $\sqrt{AVE} = 0.86$, $\alpha = 0.86$, largest correlation = 0.75)

- *SSA: Perceived Trust (T)*: 4 items (T1, T2, T3, T4). ($AVE = 0.76$, $\sqrt{AVE} = 0.86$, $\alpha = 0.91$, largest correlation = 0.85)
- *SSA: Perceived Transparency (P)*: 3 items (P1, P2, P3). ($AVE = 0.68$, $\sqrt{AVE} = 0.82$, $\alpha = 0.81$, largest correlation = 0.73)
- *EXP: Perceived Satisfaction (S)*: 4 items (S1, S2, S3, S4). ($AVE = 0.81$, $\sqrt{AVE} = 0.90$, $\alpha = 0.91$, largest correlation = 0.88)

I then built a structural equation model (SEM) for analyzing the UX concepts and the directionality of causal effects. I followed the framework recommended in [71]. I iteratively tested and removed the constructs in the SEM model until I found that the model is stable and fitted. After the analysis, I kept only three constructs: *Perceived Recommendation Quality (Q)*, Perceived Transparency (P), and Perceived Satisfaction (S) in the model. The model fit the statistics of $\chi^2(200) = 227.29$, $p < 0.001$, $RMSEA = 0.054$, $90\%CI : [0.040, 0.067]$, $CFI = 0.99$, $TLI = 1.00$, which indicates an effective fitting model.

The built model is consistent with findings by [9, 70] that user satisfaction was mediated through the constructs of perceived understandability (in my case, the construct of perceived transparency) and perceived recommendation quality. I believe the selected constructs are essential in my user experiment design, i.e., users performing information-seeking tasks through the recommender interfaces. At the same time, since the user tasks were not designed to promote recommendation diversity, gain user trust, or let the user make the decision faster (effectiveness), it is not surprising to see that the user experience was not mediated through these constructs. However, it is surprising to see the construct of perceived control had to be removed as well. The results indicated that in my study, user satisfaction was not mediated through the user perception of control; instead, the user satisfaction steams from perceived transparency and perceived recommendation quality.

The model shows that the controllability and explainability manipulations each have an independent positive effect on *perceived transparency* of the system. The controllable interface is more transparent than interfaces with no control; however, explainability contributes more to the perceived transparency than controllability. *Perceived transparency* is

in turn contributes to the *perceived quality* of the recommendations. The *perceived quality* finally determines participants' satisfaction with the system. I also found the controllability manipulations have positive effects on the factor of *perceived quality*, but the explainability manipulation has a surprisingly negative effect on the factor of *perceived quality*.

I expanded the core model by adding additional variables: user activity and user characteristics. I found three positive/negative effects of personal characteristics (PCs) on the factor of *perceived transparency*. The finding revealed a higher transparency perception if the participants have lower trust in technology (PC1), have a higher score of scholar expertise (PC2), have a higher acceptance of diversity (of recommendation) (PC3), or have higher privacy concern (SC1). A higher score of acceptance of diversity also has a direct positive effect on the factor of *perceived quality*. Interestingly, I found that *system familiarity* has a negative effect on user satisfaction, which indicates that experts may be harder to please. In the variables of user activity, I found that the need to explore more scholars in detail (more clicks on scholar name) has a negative effect on the factor of *perceived quality*. At the same time, as the participants work with the system longer (spent more time in solving the tasks), their perception of system transparency increases.

9.5 SUMMARY AND CONCLUSIONS

In this paper, I presented a large-scale human subject experiment (N=50) that assessed the impact of controllable and explainable recommender interfaces in a hybrid social recommender system. For the purpose of the study, I augmented a hybrid recommender system RelevanceTuner+, which offered a user-controllable fusing or recommendation sources through sliders with a total of four explainable source recommendation models (publication, topic, co-authorship and interest similarities). I then conducted a controlled user study using four-year proceeding data of UMAP conferences. To examine separate and compound impact of controllability and explainability, I used four conditions: Baseline interface (BASE) with both sliders and access to explanations disabled; Controllable interface (CONT) with slider enabled; Explainable interface (EXPL) with explanation access enabled; Full interface

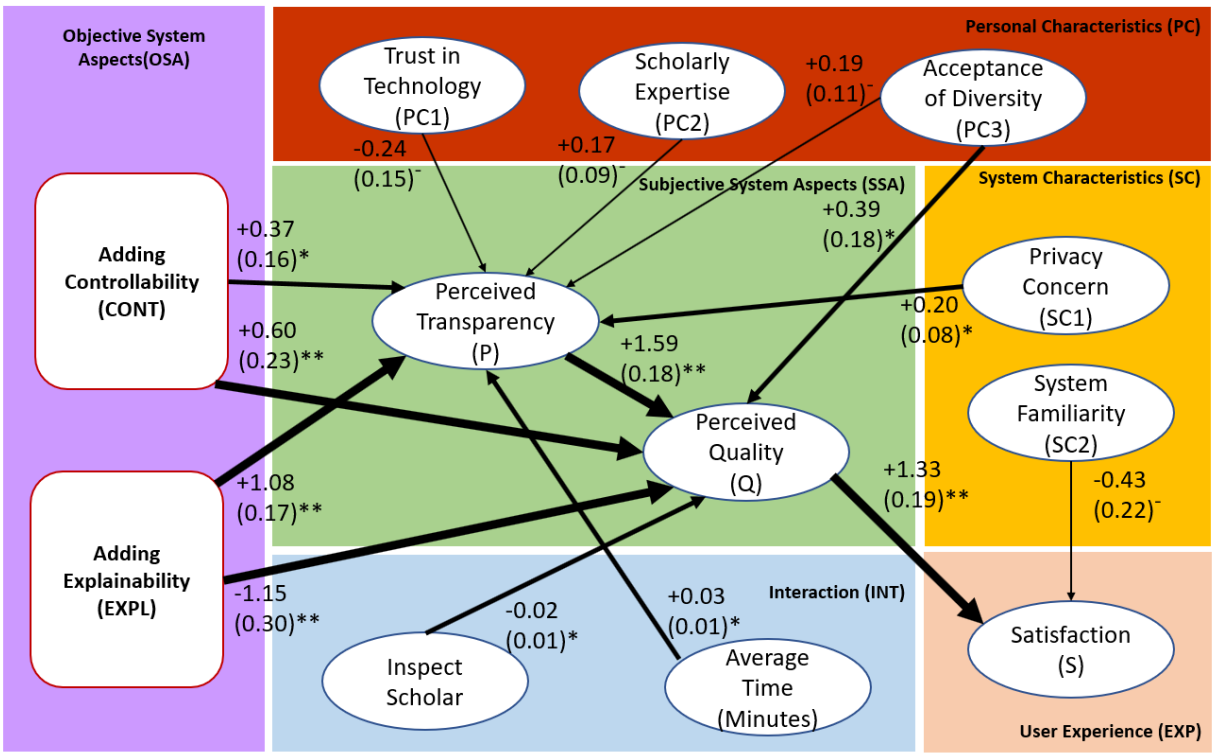


Figure 40: The structural equation model of study 6. The number (thickness) on the arrows represents the coefficients and standard error of the effect. Significance: $**p < 0.01$, $*p < 0.05$, $-p < 0.1$.

(FULL) with both slider and explanation access enabled. I followed the within-subject design and assigned the participants two scenario-based tasks, one relevance-oriented and one diversity-orientated. I used a total of 7 constructs with 19 questions as subjective measures. The objective measures were captured by analyzing user activity log and applying traditional recommendation quality metrics.

The user action analysis demonstrated that the participants extensively adopted the control and explanation functionality provided by my interfaces. I also observed that in their work with sliders, the users *boosted* the relevance weighting in the relevance-oriented task (Task 1) and *decreased* the weighting in the diverse-oriented task (Task 2) as expected by the nature of the tasks.

The analysis of recommendation quality using traditional ranking-focused metrics demonstrated that in the diversity-oriented task, the availability of controllable fusion increased the recommendation quality, that is, the users adopted more recommendations in the higher-ranked items. In contrast, providing explanation function leads the users to click more recommendations beyond top-ranked items, decreasing all ranking-based metrics. Interestingly, the analysis of alpha nDCG demonstrated that picking more items in the lower ranks may not guarantee the high recommendation diversity, which was determined by the dissimilarity between the recommending items.

The results of the subjective feedback analysis demonstrated that the *FULL* condition is outperformed all other conditions, which answers RQ1. The analysis shows evidence that the provision of the controllable and explainable user interface has a significant positive effect on user perception. Moreover, the effect of providing controllable fusion and reasoning transparency is additive. The *FULL* is outperformed than the single enhanced condition, i.e., *CONT* and *EXPL*. Another interesting finding indicates the academic background affects user perception. I found the participants who are Master students are a more "easy-to-please" group. Their subjective feedback score is significantly higher than Ph.D. students.

It is interesting to stress that traditional recommendation quality metrics and user feedback analysis are seemingly opposing conclusions on the effect of explainability. While controllability had a clear positive impact on user perception of the system, it also caused users to select relevant items in lower ranks. It is quite natural to expect that in the presence of explanations, the user has more supporting information for making their choice than with the ranking alone and would be willing to explore lower-ranked items with increased satisfaction. However, it is a clear sign that traditional ranking-focused metrics have to be applied with more caution to interactive recommender systems. In the absence of user-contributed data, the ranking-based analysis might not be sufficient to determine a better design option.

In the section of the structural equation model (SEM) analysis, I used seven subjective, three personal characteristics and three system-specific characteristic constructs. I firstly ran a confirmatory factor analysis (CFA) to confirm the convergent validity and convergent validity of constructs. I then included only the valid constructs in the SEM analysis. The results indicate that both controllability (*CONT*) and explainability (*EXPL*) positively con-

tributed to the construct of *Perceived Transparency (P)*, although the *EXPL* condition had stronger effect than *CONT* condition. However, the effect of controllability and explainability on *Perceived Quality (Q)* was the opposite. The *CONT* condition showed positive contribution but *EXPL* condition was negatively contributed to the construct.

I found the best user experience happened when both controllability and explainability were provided (*FULL*). However, I also found the interaction between the two functions. Based on the score of *Perceived Control (C)*, I found the controllability was additive when the controllable sliders were provided. That is, although explanation function contributed to *Perceived Control (C)*, but the score can be strengthened by the controllable sliders. In contrast, I found the explanation function contributed to the score of *Perceived Transparency (P)* more than the controllable slider. That is, although the slider contributed to gain recommender system transparency, the user perception can be improved by providing extra explanations.

This paper confirmed some findings of previous studies while also offering new findings and providing a deeper analysis. Overall, the results support my belief in the effectiveness of the controllable and explainable user interfaces. Moreover, I found the effects of the two explored enhancements are additive, which means the best user experience happens in the *FULL* condition.

10.0 CONCLUSIONS

10.1 SUMMARY AND CONTRIBUTIONS

In this dissertation, I explored the value of controllability and explainability in a hybrid social recommender system. I systematically demonstrated the value through a set of exploration, empirical studies, data-driven analysis, as well as statistical testing. I proposed an effective user controllable interface in a hybrid social recommender system, which had been evaluated through a pre-study, presented in chapter 4. In the study, I reported my exploration of two sets of controllable interfaces: *RANK vs. Scatter* and *Scatter vs. Relevance Tuner*. The study result supported *Relevance Tuner* is a useful interface that the users have extensively adopted as well as perceive higher user perception of system control. I conducted a further data analysis to show how do the users adopt *Relevance Tuner* in real-world information-seeking tasks. Based on the experiment, and I choose *Relevance Tuner* as the first core interface.

I then proposed user explainable interfaces in a hybrid social recommender system through study 1, 2, and 3, presented in chapter 5 and 6. I conducted a stage-based participatory study that iteratively designed and evaluated the explanation interfaces for a hybrid social recommender system. In study 1, I conducted a user study to identify the *Target Mental Model*, that is, to identify the key components of the recommendation model that the users might want to be explainable in the user interface. I firstly identify 11 factors from the card-sorting task. Based on the result of study 1, I proposed a total of 25 interfaces for five recommendation models. These interfaces were evaluated trough study 2 for determining the top-rated and second-rated designs. I then implemented the top-rated interfaces and continued with study 3 for the evaluation. The result showed the effectiveness of three ex-

planation interfaces: *E2-4 Topical Radar*, *E4-4 Venn Tags* and *E5-3 Navigation Style Map*. However, the *E1-4 Venn Word Cloud* and *E3-3 ForceAtlas2* had a low user performance that required a revision.

In study 4, I conducted an online study with real conference proceeding data. Based on the finding of study 4 (chapter 7), I confirmed the fusion transparent and reasoning transparency, i.e., the controllable and explainable interface could be combined in an interactive social recommender system. The experiment results showed a trade-off pattern between the system controllability and explainability. I then conducted the study 5 (chapter 8) for the second round of evaluation through a crowd-sourced online study. Two revised explanation interfaces, *E1-2 Two-Way Bar Chart* and *E3-3 Strength Graph* were porpoised and evaluated. I identified the effective design though this experiment.

In study 6 (chapter 9), I conducted a large-scale human subject experiment to assess the impact of controllable and explainable recommendation interfaces in a hybrid social recommender system. I proposed *Relevance Tuner+* that offered an integrated of user-controllable and explainable interfaces. The goal of this study is to confirm the interface's effectiveness as well as further analyzed the interaction between controllability and explainability. I found the best user experience happened in the *full* condition, that is, when both controllable and explainable interface was provided. Through the action analysis, recommendation quality analysis, and user feedback analysis, I showed a different pattern in terms of how do the enhanced controllability and explainability affect the user perception, user experience, and user engagement with a hybrid social recommender system.

I found there were a potential behavior and perception difference between users who have a different level of expertise. For example, in study 6, I conducted an analysis to inspect the cohort difference based on participants' academic background, i.e., doctoral and master students. I found a significant difference between the two groups in user perception analysis. In general, the master students had higher user perception scores than Ph.D. students. It provided evidence to support the correlation between user expertise and user perception, which was also discussed in the work of [115]. The finding led to a design implication that the interface design and evaluation should take user expertise into account. For example, less-expertise users may appreciate an easy-to-use interface than the professional data analyst.

It is also worth considering different approaches to evaluate the interfaces, e.g., to test the interface with less-expertise users through random web users (A/B testing) or high-expertise users through a lab-controllable study with multiple information-seeking tasks.

This analysis helps to answer the *[RQ1] How do the enhanced controllability and explainability affect the user perception, user experience, and user engagement with a hybrid social recommender system?*. I find the controllability and explainability effects are additive, which answered the *[RQ2] Is there an interaction effect between controllability and explainability in a hybrid social recommender system?* as well as indicated the best user experience happened in both recommendation fusion and reasoning were transparent. I confirmed the interaction effects between controllability and explainability in a hybrid social recommender system. I found the controllability was additive when the controllable sliders were provided. The user perception of control can be strengthened by the controllable sliders. I further found the explanation interface contributed to the user perception of system transparency more than the controllable sliders. It supported the extra layer of system transparency was also additive.

10.2 DISCUSSION

Transferability: I made several contributions that can be transferred to another domain. First, I proposed two user-controllable interfaces for the hybrid social recommendation. Second, I distilled best designs for several kinds of similarity-based explanations, using participatory design approaches. I discussed the guidelines and factors of combining recommendations and explanations. Third, I designed approaches for constructing and assessing visual explanations through the field, online, and lab-controlled studies. All these contributions have the potential to transfer to other domains.

1. *User-controllable Interfaces:* My proposed user-controllable interfaces can be adopted in different domains with content-based recommendations. For example, the *relevance tuner* has been extended to the application of paper recommendation [109], which empowers users to explore academic papers through multiple relevance scores. Another example

has lied in music recommendation that provides a user-controllable interface on multiple dimensions of music preference [93]. There are some possible domains, e.g., job recommendation [49] for comparing different candidates, course recommendation for choosing between courses [46] and health recommendation for selecting clinical treatments [112], etc.

2. *Similarity-based Explanations:* Although the design of the explanation interface is context-dependent, my works can provide insights on bringing transparency to such domains and applications. Specifically, I proposed explanation interfaces across recommendation models using different data mining approaches, i.e., text mining, clustering, graph mining, set relations, and spatial data mining. All these approaches are widely adopted in many different recommender systems or AI-driven appellations. My works provide an attempt to bring transparency to these systems. For example, text analysis for tourism recommendation [86], clustering algorithm for medical diagnosis[123], relationship recommendation for social network [16], bipartite structure recommendation [143] and spatial item recommender system [145].
3. *Evaluation Approaches:* I designed several evaluation approaches to construct and assess the proposed user interfaces, a mixture of quantitative and qualitative approaches [74]. All these attempts showed a boarder view of user experience from a different perspectives. I presented the human subject experiments from different groups of users (e.g., the hyper-local users, online system users, online crowd workers, and lab participants) and from different study designs (e.g., participatory design, between-subject design, within-subject design, and semi-interview). All these experiments can be used to answer different research questions as well as present different scientific findings. The experience can be transferred to a different stage of designing controllable and explainable AI systems. For example, co-design the system across different stakeholders [59], user-centric evaluation [87] and cross-domain personality modeling [47].

Future Works: The newly initiated European Union's General Data Protection Regulation (GDPR) required the owner of any data-driven application to maintain a "right to the explanation" of algorithmic decisions [32], which urged to gain *transparency* in all existing intelligent systems. It was also important to protect users' privacy by increasing the inter-

pretability of AI-driven applications or models. My work aims to provide an exploration of how to enhance system transparency through controllable and explainable user interfaces. The similarity-based explanation is based on revealing the *self-relevance* of the users. However, it is worth considering if adding system transparency will further violate the privacy guideline. For example, in a similarity-based explanation, my solution presents the *similarity* between two users, which may expose sensitive data that the encounter party does not want to share. A privacy tool may be able to control the data that can be used in the explanation, but the interface can only provide a limited explanation if only a few data are accessible. It worth a further exploration of adding privacy controls in the interface design.

There are still many open research questions in this field; for example, how can we increase the *reproducibility* of the proposed prototype user interfaces? In some cases, the user interfaces were evaluated in different experiment settings that may influence the findings and design implications. For example, in this dissertation, based on the participatory design process, I found the *Venn Diagram* was outperformed than the text-based explanation. However, in the work of Kouki et al. [76], the researchers reported contradicted findings, i.e., the text-based explanation was perceived better than the visualized Venn diagram. The inconsistency pointed out a challenge in the stream of research, that is, the issue of reproducibility. I found the two studies were conducted in different experimental designs and contexts. In this dissertation, I conducted a human subject lab-controlled study with a context of social recommendation. The participants came from college and university. The work of Kouki et al. [76] ran their experiment in a crowd-sourced platform with a context of music recommendation. It provides an explanation of why the findings are a contradiction to each other; however, further exploration will be required to confirm the reproducibility of the proposed user interface designs.

Limitations: I notice there are some limitations to the works of this dissertation. First, my works are a focus on similarity-based recommendations that are naturally presenting in multi-relevance structure. That is, my approaches may not apply in other recommendation models, such as correlation similarity by matrix factorization or deep neural network [23]. Second, while a controlled user study with 50 subjects (study 6) is considered as a large scale, some observed trends might not be able to reach significance due to the scale. Specifically,

a larger number of subjects would be desirable to build a reliable structural equation model of the process. Third, I also noticed that some participants, especially doctoral students, maybe highly specific in their research interests. In this situation, UMAP conference data used in my study may not offer a sufficient fit to the research interests of all participants. A better match between participants' general interests and the dataset might bring more reliable results.

BIBLIOGRAPHY

[1] H. Abdollahpouri, G. Adomavicius, R. Burke, I. Guy, D. Jannach, T. Kamishima, J. Krasnodebski, and L. Pizzato. Beyond personalization: Research directions in multistakeholder recommendation. *arXiv preprint arXiv:1905.01986*, 2019.

[2] J.-w. Ahn, P. Brusilovsky, J. Grady, D. He, and R. Florian. Semantic annotation based exploratory search for information analysts. *Information processing & management*, 46(4):383–402, 2010.

[3] J.-w. Ahn, P. Brusilovsky, D. He, J. Grady, and Q. Li. Personalized web exploration with task models. In *the 17th international conference on World Wide Web, WWW '08*, pages 1–10. ACM, 2008.

[4] S. Amershi, D. Weld, M. Vorvoreanu, A. Fourney, B. Nushi, P. Collisson, J. Suh, S. Iqbal, P. N. Bennett, K. Inkpen, et al. Guidelines for human-ai interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, page 3. ACM, 2019.

[5] S. M. Bailey, J. A. Wei, C. Wang, D. Parra, and P. Brusilovsky. Cnvis: A web-based visual analytics tool for exploring conference navigator data. *Electronic Imaging*, 2018(1):1–11, 2018.

[6] A. Bellogín, I. Cantador, F. Díez, P. Castells, and E. Chavarriaga. An empirical comparison of social, collaborative filtering, and hybrid recommenders. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 4(1):14, 2013.

[7] T. Bogers. Tag-based recommendation. In *Social Information Access*, pages 441–479. Springer, 2018.

[8] D. Bollen, B. P. Knijnenburg, M. C. Willemsen, and M. Graus. Understanding choice overload in recommender systems. In *Proceedings of the fourth ACM conference on Recommender systems*, pages 63–70. ACM, 2010.

[9] S. Bostandjiev, J. O'Donovan, and T. Höllerer. Tasteweights: a visual interactive hybrid recommender system. In *Proceedings of the sixth ACM conference on Recommender systems*, pages 35–42. ACM, 2012.

[10] P. Brusilovsky, J. S. Oh, C. López, D. Parra, and W. Jeng. Linking information and people in a social system for academic conferences. *New Review of Hypermedia and Multimedia*, pages 1–31, 2016.

[11] B. G. Buchanan, E. H. Shortliffe, et al. *Rule-based expert systems*, volume 3. Addison-Wesley Reading, MA, 1984.

[12] R. Burke. Hybrid web recommender systems. In P. Brusilovsky, A. Kobsa, and W. Neidl, editors, *The Adaptive Web: Methods and Strategies of Web Personalization*, volume 4321 of *Lecture Notes in Computer Science*, pages 377–408. Springer-Verlag, Berlin Heidelberg New York, 2007.

[13] B. Cardoso, G. Sedrakyan, F. Gutiérrez, D. Parra, P. Brusilovsky, and K. Verbert. Intersectionexplorer, a multi-perspective approach for exploring recommendations. *International Journal of Human-Computer Studies*, 2018.

[14] S. Castagnos, A. Brun, and A. Boyer. When diversity is needed... but not expected! In *International Conference on Advances in Information Mining and Management*, pages 44–50. IARIA XPS Press, 2013.

[15] Ò. Celma Herrada. *Music recommendation and discovery in the long tail*. PhD thesis, Universitat Pompeu Fabra, 2009.

[16] P. Chamoso, A. Rivas, S. Rodríguez, and J. Bajo. Relationship recommender system in a business and employment-oriented social network. *Information sciences*, 433:204–220, 2018.

[17] J. Chen, W. Geyer, C. Dugan, M. Muller, and I. Guy. Make new friends, but keep the old: recommending people on social networking sites. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 201–210. ACM, 2009.

[18] L. Chen and P. Pu. Trust building in recommender agents. In *Proceedings of the Workshop on Web Personalization, Recommender Systems and Intelligent User Interfaces at the 2nd International Conference on E-Business and Telecommunication Networks*, pages 135–145, 2005.

[19] L. Chen and F. Wang. Explaining recommendations based on feature sentiments in product reviews. In *Proceedings of the 22Nd International Conference on Intelligent User Interfaces*, IUI '17, pages 17–28, New York, NY, USA, 2017. ACM.

[20] H.-F. Cheng, R. Wang, Z. Zhang, F. O'Connell, T. Gray, F. M. Harper, and H. Zhu. Explaining decision-making algorithms through ui: Strategies to help non-expert stakeholders. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, page 559. ACM, 2019.

[21] G. Chowdhury. *Introduction to modern information retrieval*. Facet publishing, 2010.

[22] C. L. Clarke, M. Kolla, G. V. Cormack, O. Vechtomova, A. Ashkan, S. Büttcher, and I. MacKinnon. Novelty and diversity in information retrieval evaluation. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 659–666. ACM, 2008.

[23] P. Covington, J. Adams, and E. Sargin. Deep neural networks for youtube recommendations. In *Proceedings of the 10th ACM conference on recommender systems*, pages 191–198. ACM, 2016.

[24] H. Cramer, V. Evers, S. Ramlal, M. Van Someren, L. Rutledge, N. Stash, L. Aroyo, and B. Wielinga. The effects of transparency on trust in and acceptance of a content-based art recommender. *User Modeling and User-Adapted Interaction*, 18(5):455, 2008.

[25] N. Craswell. Mean reciprocal rank. *Encyclopedia of Database Systems*, pages 1703–1703, 2009.

[26] P. Cremonesi, F. Garzotto, and R. Turrin. Investigating the persuasion potential of recommender systems from a quality perspective: An empirical study. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 2(2):11, 2012.

[27] T. N. Dang and L. Wilkinson. Scagexplorer: Exploring scatterplots by their scagnostics. In *Visualization Symposium (PacificVis), 2014 IEEE Pacific*, pages 73–80. IEEE, 2014.

[28] S. Deeswe and R. Kosala. An integrated search interface with 3d visualization. *Procedia Computer Science*, 59:483–492, 2015.

[29] C. di Sciascio, P. Brusilovsky, and E. Veas. A study on user-controllable social exploratory search. In *23rd International Conference on Intelligent User Interfaces*, pages 353–364. ACM, 2018.

[30] C. di Sciascio, V. Sabol, and E. E. Veas. Rank as you go: User-driven exploration of search results. In *Proceedings of the 21st International Conference on Intelligent User Interfaces*, pages 118–129. ACM, 2016.

[31] J. Du, J. Jiang, D. Song, and L. Liao. Topic modeling with document relative similarities. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*, 2015.

[32] M. Eiband, H. Schneider, M. Bilandzic, J. Fazekas-Con, M. Haug, and H. Hussmann. Bringing transparency design into practice. In *23rd International Conference on Intelligent User Interfaces*, pages 211–223. ACM, 2018.

[33] M. D. Ekstrand, F. M. Harper, M. C. Willemsen, and J. A. Konstan. User perception of differences in recommender algorithms. In *Proceedings of the 8th ACM Conference on Recommender systems*, pages 161–168. ACM, 2014.

[34] M. D. Ekstrand, D. Kluver, F. M. Harper, and J. A. Konstan. Letting users choose recommender algorithms: An experimental study. In *Proceedings of the 9th ACM Conference on Recommender Systems*, pages 11–18. ACM, 2015.

[35] N. B. Ellison, C. Steinfield, and C. Lampe. Connection strategies: Social capital implications of facebook-enabled communication practices. *New media & society*, 13(6):873–892, 2011.

[36] S. Faridani, E. Bitton, K. Ryokai, and K. Goldberg. Opinion space: a scalable tool for browsing online comments. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1175–1184. ACM, 2010.

[37] G. Friedrich and M. Zanker. A taxonomy for generating explanations in recommender systems. *AI Magazine*, 32(3):90–98, 2011.

[38] A. Garnett, G. Lee, and J. Illes. Publication trends in neuroimaging of minimally conscious states. *PeerJ*, 1:e155, 2013.

[39] M. Ge, F. Gedikli, and D. Jannach. Placing high-diversity items in top-n recommendation lists. In *Workshop chairs*, page 65. Citeseer, 2011.

[40] F. Gedikli, M. Ge, and D. Jannach. Understanding recommendations by reading the clouds. In *International Conference on Electronic Commerce and Web Technologies*, pages 196–208. Springer, 2011.

[41] F. Gedikli, D. Jannach, and M. Ge. How should i explain? a comparison of different explanation types for recommender systems. *International Journal of Human-Computer Studies*, 72(4):367–382, 2014.

[42] D. Glowacka, T. Ruotsalo, K. Konuyshkova, S. Kaski, G. Jacucci, et al. Directing exploratory search: Reinforcement learning from user interactions with keywords. In *Proceedings of the 2013 international conference on Intelligent user interfaces*, pages 117–128. ACM, 2013.

[43] W. H. Gomaa and A. A. Fahmy. A survey of text similarity approaches. *International Journal of Computer Applications*, 68(13):13–18, 2013.

[44] E. Graells-Garrido, M. Lalmas, and R. Baeza-Yates. Data portraits and intermediary topics: Encouraging exploration of politically diverse profiles. In *Proceedings of the 21st International Conference on Intelligent User Interfaces*, pages 228–240. ACM, 2016.

[45] B. Gretarsson, J. O'Donovan, S. Bostandjiev, C. Hall, and T. Höllerer. Smallworlds: visualizing social recommendations. In *Computer Graphics Forum*, pages 833–842. Wiley Online Library, 2010.

[46] J. Guerra-Hollstein, J. Barria-Pineda, C. D. Schunn, S. Bull, and P. Brusilovsky. Fine-grained open learner models: Complexity versus support. In *Proceedings of the 25th*

Conference on User Modeling, Adaptation and Personalization, pages 41–49. ACM, 2017.

[47] S. C. Guntuku, S. Roy, and L. Weisi. Personality modeling based image recommendation. In *International Conference on Multimedia Modeling*, pages 171–182. Springer, 2015.

[48] P. Gupta, A. Goel, J. Lin, A. Sharma, D. Wang, and R. Zadeh. Wtf: The who to follow service at twitter. In *Proceedings of the 22nd international conference on World Wide Web*, pages 505–514. ACM, 2013.

[49] F. Gutiérrez, S. Charleer, R. De Croon, N. N. Htun, G. Goetschalckx, and K. Verbert. Explaining and exploring job recommendations: a user-driven approach for interacting with knowledge-based job recommender systems. In *Proceedings of the 13th ACM Conference on Recommender Systems*, pages 60–68. ACM, 2019.

[50] I. Guy. Social recommender systems. In *Recommender Systems Handbook*, pages 511–543. Springer, 2015.

[51] I. Guy. People recommendation on social media. In *Social Information Access*, pages 570–623. Springer, 2018.

[52] I. Guy, I. Ronen, and E. Wilcox. Do you know?: recommending people to invite into your social network. In *Proceedings of the 14th international conference on Intelligent user interfaces*, pages 77–86. ACM, 2009.

[53] I. Guy, S. Ur, I. Ronen, A. Perer, and M. Jacovi. Do you want to know?: recommending strangers in the enterprise. In *Proceedings of the ACM 2011 conference on Computer supported cooperative work*, pages 285–294. ACM, 2011.

[54] S. Han, D. He, J. Jiang, and Z. Yue. Supporting exploratory people search: a study of factor transparency and user control. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*, pages 449–458. ACM, 2013.

[55] F. M. Harper, F. Xu, H. Kaur, K. Condiff, S. Chang, and L. Terveen. Putting users in control of their recommendations. In *Proceedings of the 9th ACM Conference on Recommender Systems*, pages 3–10. ACM, 2015.

[56] C. He, D. Parra, and K. Verbert. Interactive recommender systems: a survey of the state of the art and future research challenges and opportunities. *Expert Systems with Applications*, 56:9–27, 2016.

[57] R. Heckel, M. Vlachos, T. Parnell, and C. Dünner. Scalable and interpretable product recommendations via overlapping co-clustering. In *Data Engineering (ICDE), 2017 IEEE 33rd International Conference on*, pages 1033–1044. IEEE, 2017.

[58] J. L. Herlocker, J. A. Konstan, and J. Riedl. Explaining collaborative filtering recommendations. In *Proceedings of the 2000 ACM conference on Computer supported cooperative work*, pages 241–250. ACM, 2000.

[59] I. Holeman, E. Blake, M. Densmore, M. Molapo, F. Ssozi, E. Goodman, I. Medhi Thies, and S. Wyche. Co-design across borders special interest group. In *Proceedings of the 2017 CHI conference extended abstracts on human factors in computing systems*, pages 1318–1321. ACM, 2017.

[60] R. Hu and P. Pu. Helping users perceive recommendation diversity. In *DiveRS@ RecSys*, pages 43–50, 2011.

[61] G. Inc. Google maps directions api. https://developers.google.com/maps/documentation/directions/intro. Web Accessed: 2019-01-02.

[62] D. Jannach, S. Naveed, and M. Jugovac. User control in recommender systems: Overview and interaction challenges. In *International Conference on Electronic Commerce and Web Technologies*, pages 21–33. Springer, 2016.

[63] M. Kaminskas and D. Bridge. Diversity, serendipity, novelty, and coverage: A survey and empirical analysis of beyond-accuracy objectives in recommender systems. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 7(1):2, 2016.

[64] M. Keane, M. O'Brien, and B. Smyth. Are people biased in their use of search engines? *Commun. ACM*, 51(2):49–52, 2008.

[65] B. Kim. *Interactive and interpretable machine learning models for human machine collaboration*. PhD thesis, Massachusetts Institute of Technology, 2015.

[66] H. Kim, J. Choo, H. Park, and A. Endert. Interaxis: Steering scatterplot axes via observation-level interaction. *IEEE transactions on visualization and computer graphics*, 22(1):131–140, 2016.

[67] K. Klouche, T. Ruotsalo, D. Cabral, S. Andolina, A. Bellucci, and G. Jacucci. Designing for exploratory search on touch devices. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 4189–4198. ACM, 2015.

[68] K. Klouche, T. Ruotsalo, L. Micallef, S. Andolina, and G. Jacucci. Visual re-ranking for multi-aspect information retrieval. In *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval*, pages 57–66. ACM, 2017.

[69] D. Kluver, M. D. Ekstrand, and J. A. Konstan. Rating-based collaborative filtering: algorithms and evaluation. In *Social Information Access*, pages 344–390. Springer, 2018.

[70] B. P. Knijnenburg, S. Bostandjiev, J. O'Donovan, and A. Kobsa. Inspectability and control in social recommenders. In *Proceedings of the sixth ACM conference on Recommender systems*, pages 43–50. ACM, 2012.

[71] B. P. Knijnenburg and M. C. Willemsen. Evaluating recommender systems with user experiments. In *Recommender Systems Handbook*, pages 309–352. Springer, 2015.

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

[73] A. Kobsa, J. Koenemann, and W. Pohl. Personalised hypermedia presentation techniques for improving online customer relationships. *The knowledge engineering review*, 16(2):111–155, 2001.

[74] Y. Kou, C. M. Gray, A. L. Toombs, and R. S. Adams. Understanding social roles in an online community of volatile practice: A study of user experience practitioners on reddit. *ACM Transactions on Social Computing*, 1(4):17, 2018.

[75] P. Kouki, J. Schaffer, J. Pujara, J. O'Donovan, and L. Getoor. User preferences for hybrid explanations. In *Proceedings of the Eleventh ACM Conference on Recommender Systems*, pages 84–88. ACM, 2017.

[76] P. Kouki, J. Schaffer, J. Pujara, J. O'Donovan, and L. Getoor. Personalized explanations for hybrid recommender systems. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*, pages 379–390. ACM, 2019.

[77] T. Kulesza, M. Burnett, W.-K. Wong, and S. Stumpf. Principles of explanatory debugging to personalize interactive machine learning. In *Proceedings of the 20th international conference on intelligent user interfaces*, pages 126–137. ACM, 2015.

[78] T. Kulesza, S. Stumpf, M. Burnett, S. Yang, I. Kwan, and W.-K. Wong. Too much, too little, or just right? ways explanations impact end users' mental models. In *Visual Languages and Human-Centric Computing (VL/HCC), 2013 IEEE Symposium on*, pages 3–10. IEEE, 2013.

[79] J. Kunkel, B. Loepp, and J. Ziegler. A 3d item space visualization for presenting and manipulating user preferences in collaborative filtering. In *Proceedings of the 22nd International Conference on Intelligent User Interfaces*, pages 3–15. ACM, 2017.

[80] Lawrence. Customize d3plus network style. https://codepen.io/choznerol/pen/evaYyv. Web Accessed: 2019-01-02.

[81] D. Lee and P. Brusilovsky. How to measure information similarity in online social networks: A case study of citeulike. *Information Sciences*, 418-419:46–60, 2017.

[82] D. Lee and P. Brusilovsky. Recommendations based on social links. In *Social Information Access*, pages 391–440. Springer, 2018.

[83] Q. V. Liao and W.-T. Fu. Beyond the filter bubble: interactive effects of perceived threat and topic involvement on selective exposure to information. In *Proceedings of the*

SIGCHI conference on human factors in computing systems, pages 2359–2368. ACM, 2013.

[84] B. Y. Lim and A. K. Dey. Assessing demand for intelligibility in context-aware applications. In *Proceedings of the 11th international conference on Ubiquitous computing*, pages 195–204. ACM, 2009.

[85] B. Liu. *Better than PageRank: Hitting Time as a Reputation Mechanism*. PhD thesis, 2014.

[86] S. Loh, F. Lorenzi, R. Saldaña, and D. Licthnow. A tourism recommender system based on collaboration and text analysis. *Information Technology & Tourism*, 6(3):157–165, 2003.

[87] M. Ludewig and D. Jannach. User-centric evaluation of session-based recommendations for an automated radio station. In *Proceedings of the 13th ACM Conference on Recommender Systems*, pages 516–520. ACM, 2019.

[88] C. Manning, P. Raghavan, and H. Schütze. Introduction to information retrieval. *Natural Language Engineering*, 16(1):100–103, 2010.

[89] C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to information retrieval*, volume 1. Cambridge university press Cambridge, 2008.

[90] J. McAuley and J. Leskovec. Hidden factors and hidden topics: understanding rating dimensions with review text. In *Proceedings of the 7th ACM conference on Recommender systems*, pages 165–172. ACM, 2013.

[91] S. McNee, S. Lam, J. Konstan, and J. Riedl. Interfaces for eliciting new user preferences in recommender systems. *User Modeling 2003*, pages 148–148, 2003.

[92] A. W. Meade and S. B. Craig. Identifying careless responses in survey data. *Psychological methods*, 17(3):437, 2012.

[93] M. Millecamp, N. N. Htun, C. Conati, and K. Verbert. To explain or not to explain: the effects of personal characteristics when explaining music recommendations. In *IUI*, pages 397–407, 2019.

[94] J. Moody and D. H. Glass. A novel classification framework for evaluating individual and aggregate diversity in top-n recommendations. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 7(3):42, 2016.

[95] M. E. Newman. Clustering and preferential attachment in growing networks. *Physical Review E*, 64(2):025102, 2001.

[96] T. T. Nguyen, P.-M. Hui, F. M. Harper, L. Terveen, and J. A. Konstan. Exploring the filter bubble: the effect of using recommender systems on content diversity. In

Proceedings of the 23rd international conference on World wide web, pages 677–686. ACM, 2014.

[97] J. O'Donovan, B. Gretarsson, S. Bostandjiev, T. Hollerer, and B. Smyth. A visual interface for social information filtering. In *Computational Science and Engineering, 2009. CSE'09. International Conference on*, volume 4, pages 74–81. IEEE, 2009.

[98] J. O'Donovan, B. Smyth, B. Gretarsson, S. Bostandjiev, and T. Höllerer. Peerchooser: visual interactive recommendation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1085–1088. ACM, 2008.

[99] M. P. O'Mahony and B. Smyth. From opinions to recommendations. In *Social Information Access*, pages 480–509. Springer, 2018.

[100] V. Orso, T. Ruotsalo, J. Leino, L. Gamberini, and G. Jacucci. Overlaying social information: The effects on users' search and information-selection behavior. *Information Processing & Management*, 53(6):1269–1286, 2017.

[101] H. J. C. Pampın, H. Jerbi, and M. P. OMahony. Evaluating the relative performance of neighbourhood-based recommender systems. In *Proceedings of the 3rd Spanish Conference on Information Retrieval*, 2014.

[102] A. V. Pandey, J. Krause, C. Felix, J. Boy, and E. Bertini. Towards understanding human similarity perception in the analysis of large sets of scatter plots. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 3659–3669. ACM, 2016.

[103] A. Papadimitriou, P. Symeonidis, and Y. Manolopoulos. A generalized taxonomy of explanations styles for traditional and social recommender systems. *Data Mining and Knowledge Discovery*, 24(3):555–583, 2012.

[104] D. Parra and P. Brusilovsky. User-controllable personalization: A case study with setfusion. *International Journal of Human-Computer Studies*, 78:43–67, 2015.

[105] D. Parra, W. Jeng, P. Brusilovsky, C. López, and S. Sahebi. Conference navigator 3: An online social conference support system. In *UMAP Workshops*, pages 1–4, 2012.

[106] P. Pu and L. Chen. Trust building with explanation interfaces. In *Proceedings of the 11th international conference on Intelligent user interfaces*, pages 93–100. ACM, 2006.

[107] P. Pu and L. Chen. Trust-inspiring explanation interfaces for recommender systems. *Knowledge-Based Systems*, 20(6):542–556, 2007.

[108] P. Pu, L. Chen, and R. Hu. A user-centric evaluation framework for recommender systems. In *Proceedings of the fifth ACM conference on Recommender systems*, pages 157–164. ACM, 2011.

[109] B. Rahdari and P. Brusilovsky. User-controlled hybrid recommendation for academic papers. In *Proceedings of the 24th International Conference on Intelligent User Interfaces: Companion*, pages 99–100. ACM, 2019.

[110] M. T. Ribeiro, A. Lacerda, A. Veloso, and N. Ziviani. Pareto-efficient hybridization for multi-objective recommender systems. In *Proceedings of the sixth ACM conference on Recommender systems*, pages 19–26. ACM, 2012.

[111] T. Ruotsalo, J. Peltonen, M. Eugster, D. Głowacka, K. Konyushkova, K. Athukorala, I. Kosunen, A. Reijonen, P. Myllymäki, G. Jacucci, et al. Directing exploratory search with interactive intent modeling. In *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management*, pages 1759–1764. ACM, 2013.

[112] H. Schäfer, S. Hors-Fraile, R. P. Karumur, A. Calero Valdez, A. Said, H. Torkamaan, T. Ulmer, and C. Trattner. Towards health (aware) recommender systems. In *Proceedings of the 2017 international conference on digital health*, pages 157–161. ACM, 2017.

[113] J. B. Schafer, J. A. Konstan, and J. Riedl. Meta-recommendation systems: user-controlled integration of diverse recommendations. In *Proceedings of the eleventh international conference on Information and knowledge management*, pages 43–51. ACM, 2002.

[114] J. Schaffer, P. Giridhar, D. Jones, T. Höllerer, T. Abdelzaher, and J. O'Donovan. Getting the message?: A study of explanation interfaces for microblog data analysis. In *Proceedings of the 20th International Conference on Intelligent User Interfaces*, pages 345–356. ACM, 2015.

[115] J. Schaffer, J. O'Donovan, J. Michaelis, A. Raglin, and T. Höllerer. I can do better than your ai: expertise and explanations. In *IUI*, pages 240–251, 2019.

[116] A. Sharma and D. Cosley. Do social explanations work?: studying and modeling the effects of social explanations in recommender systems. In *Proceedings of the 22nd international conference on World Wide Web*, pages 1133–1144. ACM, 2013.

[117] R. L. Sie, H. Drachsler, M. Bitter-Rijpkema, and P. Sloep. To whom and why should i connect? co-author recommendation based on powerful and similar peers. *Int. J. Technology Enhanced Learning*, 4(1/2):121–137, 2012.

[118] J. Silge and D. Robinson. tidytext: Text mining and analysis using tidy data principles in r. *The Journal of Open Source Software*, 1(3):37, 2016.

[119] R. Sinha and K. Swearingen. The role of transparency in recommender systems. In *CHI'02 extended abstracts on Human factors in computing systems*, pages 830–831. ACM, 2002.

[120] K. Swearingen and R. Sinha. Beyond algorithms: An hci perspective on recommender systems. In *ACM SIGIR 2001 Workshop on Recommender Systems*, 2001.

[121] R. Tarjan. Depth-first search and linear graph algorithms. *SIAM journal on computing*, 1(2):146–160, 1972.

[122] C. H. Teo, H. Nassif, D. Hill, S. Srinivasan, M. Goodman, V. Mohan, and S. Vishwanathan. Adaptive, personalized diversity for visual discovery. In *Proceedings of the 10th ACM Conference on Recommender Systems*, pages 35–38. ACM, 2016.

[123] N. D. Thanh, M. Ali, et al. A novel clustering algorithm in a neutrosophic recommender system for medical diagnosis. *Cognitive Computation*, 9(4):526–544, 2017.

[124] N. Tintarev and J. Masthoff. Designing and evaluating explanations for recommender systems. In *Recommender systems handbook*, pages 479–510. Springer, 2011.

[125] N. Tintarev and J. Masthoff. Evaluating the effectiveness of explanations for recommender systems. *User Modeling and User-Adapted Interaction*, 22(4-5):399–439, 2012.

[126] N. Tintarev and J. Masthoff. Explaining recommendations: Design and evaluation. In *Recommender Systems Handbook*, pages 353–382. Springer, 2015.

[127] C.-H. Tsai. An interactive and interpretable interface for diversity in recommender systems. In *Proceedings of the 22Nd International Conference on Intelligent User Interfaces Companion*, IUI '17 Companion, pages 225–228, New York, NY, USA, 2017. ACM.

[128] C.-H. Tsai and P. Brusilovsky. A personalized people recommender system using global search approach. *IConference 2016 Proceedings*, 2016.

[129] C.-H. Tsai and P. Brusilovsky. Enhancing recommendation diversity through a dual recommendation interface. In *Workshop on Interfaces and Human Decision Making for Recommender Systems*, 2017.

[130] C.-H. Tsai and P. Brusilovsky. Leveraging interfaces to improve recommendation diversity. In *Adjunct Publication of the 25th Conference on User Modeling, Adaptation and Personalization*, pages 65–70. ACM, 2017.

[131] C.-H. Tsai and P. Brusilovsky. Providing control and transparency in a social recommender system for academic conferences. In *Proceedings of the 25th Conference on User Modeling, Adaptation and Personalization*, pages 313–317. ACM, 2017.

[132] C.-H. Tsai and P. Brusilovsky. Beyond the ranked list: User-driven exploration and diversification of social recommendation. In *23rd International Conference on Intelligent User Interfaces*, IUI '18, pages 239–250, New York, NY, USA, 2018. ACM.

[133] C.-H. Tsai and P. Brusilovsky. Explaining social recommendations to casual users: Design principles and opportunities. In *Proceedings of the 23rd International Conference on Intelligent User Interfaces Companion*, page 59. ACM, 2018.

[134] C.-H. Tsai and P. Brusilovsky. Designing explanation interfaces for transparency and beyond. In *Workshop on Intelligent User Interfaces for Algorithmic Transparency in Emerging Technologies*, 2019.

[135] C.-H. Tsai and P. Brusilovsky. Evaluating visual explanations for similarity-based recommendations: User perception and performance. In *Proceedings of the 27th ACM Conference on User Modeling, Adaptation and Personalization*, pages 22–30. ACM, 2019.

[136] C.-H. Tsai and P. Brusilovsky. Explaining recommendations in an interactive hybrid social recommender. In *Proceedings of the 2019 Conference on Intelligent User Interface*, pages 1–12. ACM, 2019.

[137] C.-H. Tsai and P. Brusilovsky. Exploring social recommendations with visual diversity-promoting interfaces. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 10(1):5, 2019.

[138] C.-H. Tsai and Y.-R. Lin. Tracing and predicting collaboration for junior scholars. In *Proceedings of the 25th International Conference Companion on World Wide Web*, pages 375–380. International World Wide Web Conferences Steering Committee, 2016.

[139] S. Vargas and P. Castells. Rank and relevance in novelty and diversity metrics for recommender systems. In *Proceedings of the fifth ACM conference on Recommender systems*, pages 109–116. ACM, 2011.

[140] K. Verbert, D. Parra, P. Brusilovsky, and E. Duval. Visualizing recommendations to support exploration, transparency and controllability. In *Proceedings of the 2013 international conference on Intelligent user interfaces*, pages 351–362. ACM, 2013.

[141] J. Vig, S. Sen, and J. Riedl. Tagsplanations: explaining recommendations using tags. In *Proceedings of the 14th international conference on Intelligent user interfaces*, pages 47–56. ACM, 2009.

[142] J. Vig, S. Sen, and J. Riedl. The tag genome: Encoding community knowledge to support novel interaction. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 2(3):13, 2012.

[143] Q.-X. Wang, J. Li, X. Luo, J.-J. Xu, and M.-S. Shang. Effects of the bipartite structure of a network on performance of recommenders. *Physica A: Statistical Mechanics and its Applications*, 492:1257–1266, 2018.

[144] W. Wang and I. Benbasat. Recommendation agents for electronic commerce: Effects of explanation facilities on trusting beliefs. *Journal of Management Information Systems*, 23(4):217–246, 2007.

[145] W. Wang, H. Yin, S. Sadiq, L. Chen, M. Xie, and X. Zhou. Spore: A sequential personalized spatial item recommender system. In *2016 IEEE 32nd International Conference on Data Engineering (ICDE)*, pages 954–965. IEEE, 2016.

[146] M. C. Willemsen, B. P. Knijnenburg, M. P. Graus, L. C. Velter-Bremmers, and K. Fu. Using latent features diversification to reduce choice difficulty in recommendation lists. *RecSys*, 11(2011):14–20, 2011.

[147] D. Wong, S. Faridani, E. Bitton, B. Hartmann, and K. Goldberg. The diversity donut: enabling participant control over the diversity of recommended responses. In *CHI'11 Extended Abstracts on Human Factors in Computing Systems*, pages 1471–1476. ACM, 2011.

[148] Y. Wu and M. Ester. Flame: A probabilistic model combining aspect based opinion mining and collaborative filtering. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, WSDM '15, pages 199–208, New York, NY, USA, 2015. ACM.

[149] Y. Wu and M. Ester. Flame: A probabilistic model combining aspect based opinion mining and collaborative filtering. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, pages 199–208. ACM, 2015.

[150] C. Yu, L. Lakshmanan, and S. Amer-Yahia. It takes variety to make a world: diversification in recommender systems. In *Proceedings of the 12th international conference on extending database technology: Advances in database technology*, pages 368–378. ACM, 2009.

[151] C. Yu, L. V. Lakshmanan, and S. Amer-Yahia. Recommendation diversification using explanations. In *Data Engineering, 2009. ICDE'09. IEEE 25th International Conference on*, pages 1299–1302. IEEE, 2009.

[152] Y. C. Zhang, D. Ó. Séaghdha, D. Quercia, and T. Jambor. Auralist: introducing serendipity into music recommendation. In *Proceedings of the fifth ACM international conference on Web search and data mining*, pages 13–22. ACM, 2012.

[153] C.-N. Ziegler, S. M. McNee, J. A. Konstan, and G. Lausen. Improving recommendation lists through topic diversification. In *Proceedings of the 14th international conference on World Wide Web*, pages 22–32. ACM, 2005.

[154] Zoomdata. Real-time interactive zoomdata wordcloud. https://visual.ly/community/interactive-graphic/social-media/real-time-interactive-zoomdata-wordcloud. Web Accessed: 2019-01-02.