USING EXTERNAL SOURCES TO IMPROVE RESEARCH TALK RECOMMENDATION IN SMALL COMMUNITIES

by

Chirayu Wongchokprasitti

B.Eng. Computer Engineering, Chulalongkorn University, Thailand 2000

M.S. Information Science, University of Pittsburgh, 2004

Submitted to the Graduate Faculty of

School of Information Sciences in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2015

UNIVERSITY OF PITTSBURGH

SCHOOL OF INFORMATION SCIENCES

This dissertation was presented

by

Chirayu Wongchokprasitti

It was defended on

April 15, 2015

and approved by

Daqing He, PhD, Associate Professor, School of Information Sciences

Stephen Hirtle, PhD, Professor, School of Information Sciences

Jiangtao Wang, PhD, Assistant Professor, Department of Computer Science

Dissertation Advisor: Peter Brusilovsky, PhD, Professor, School of Information Sciences

USING EXTERNAL SOURCES TO IMPROVE RESEARCH TALK RECOMMENDATION IN SMALL COMMUNITIES

Chirayu Wongchokprasitti, PhD

University of Pittsburgh, 2015

In academic research communities, a typical way to spread ideas or seek for collaboration is through research talks, which might be presented at departmental colloquia or might be in given at conferences. Given a large number of research talks, with some of them happening in parallel, it becomes increasingly harder to focus on those of that are of most interest. To solve this problem, talk recommendation systems can help academics identify the most useful talks among many. This dissertation investigates methods to improve research talk recommendations, both for conference attendees and for faculty and students at a research university. More specifically, the focus of this thesis is the use of external information about user interests as a way to address the challenges of having limited data about target users. The thesis examines several kinds of external sources such as user home page, bibliography, external bookmarks, and user profiles from external information systems and explores impact of this information on the quality of talk recommendation in a general situation and in a cold-start context. For this study, the dissertation uses data from two existing talk recommendation systems, CoMeT and Conference Navigator 3, and an academic paper search system, SciNet.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

PREFACE

First and foremost, this dissertation could not have accomplished without the full support of my advisor, Professor Peter Brusilovsky. Dr Brusilovsky always provided me countless wisdoms and invaluable advices. With his patient guidance, I can complete this PhD program. I greatly appreciate the other members of my committee: Professor Stephen Hirtle, Professor Daqing He, and Professor Jiangtao Wang, who gave me insight comments and suggestions.

I am really thankful for the supports of my PAWS lab friends, Shaghayegh Sahebi, Roya Hosseini, Jennifer Yi-Ling Lin, Sharon Hsiao, Claudia Lopez, Denis Parra, Julio Daniel Guerra, Xidao Wen, and Shuguang Han. I feel very fortunate and privileged to know them. You guys are awesome! I thank all participants who dedicated their time to complete my study. Another big thank comes to Dr Siriluck Tipmongkonsilp and Dr Worasit Choochaiwattana for rescuing me so many times. I would like to thank other people who helped me along this journey.

Finally, yet the most important, I would like to thank to my late father, Prasit, my mother, Pornawee, and my sister, Teeraporn, for endless and unconditional love, support, and inspiration. Dad, finally, I made it. Without all of you guys, I cannot fulfill my dream.

1.0 INTRODUCTION

In typical research communities, whether in academia or industries, a short (typically one hour long or less) research talk is one of the most common ways to spread or seek new ideas and obtain valuable feedback. Research colleges, universities and institutes hold dozens to hundreds of research talks and series of seminars every semester. In addition, research communities and organizations usually arrange the research conferences or meetings annually. Talks feature a range of speakers, from well-known researchers and scholars to PhD students. Some talks are arranged well in advance (including conference paper presentations) while others are organized when the opportunity presents itself, within a matter of days. Talks themselves associated with a "presentation date" have been considered special items with an "expiration date" (Minkov et al., 2010). After time has passed, it has less value or even no value as a presentation recommendation to the users. Even though a number of talks were relatively large, the number of bookmarks or ratings was too small. This is a common phenomenon in social systems know as "under-contribution" (Farzan et al., 2008). Moreover, the recommender systems have little information regarding new users or users who provide little information about their interests. This problem is called "cold-start". Faced with these challenges, the recommender systems have a difficult time generating good talk recommendations.

Even though, there is external information or metadata related to talks, such as bibliography of speakers and their research colleagues and the external sources of information about users, in this dissertation, research was conducted merely on the external sources of users. Specifically, in this context, scholars and researchers usually have several external sources of information that could be used to deduce their interests, such as the personal homepage, bibliography, or the scholarly paper bookmarks made in social bookmarking systems like Mendeley or CuteULike. In this thesis, we used straightforward criteria of external source selection: (a) the sources should be easy-to-collect, (b) include textual content, and (c) contain information related to user research interests.

Our target users, scholars and researchers, usually have their official homepage to host their contact information, teaching courses, publications, and their other interests. Homepage, therefore, was selected as the first external source in this research. Even though only certain segments of homepages contain textual content related to research interests, to keep the experiment simple, a personal webpage was considered as a single information item. The second source chosen was user publication list or bibliography. Bibliography contains a series of academic papers that a user has published. With online services such as Google Scholar[1], Microsoft Academic Search[2], or Scopus[3] scientific literature database, users can be easily identified by their name and affiliation and an up-to-date collection of their publications can be downloaded. The third is a series of articles or academic papers that user have bookmarked in academic bookmarking systems. Scholars and researchers spend considerable time reading

[1] https://scholar.google.com/
[2] http://academic.research.microsoft.com/
[3] http://www.scopus.com/

scientific papers. In order to keep track of the most interesting articles, they usually store them in offline systems such as EndNote, or bookmark them in the social bookmarking systems such as CiteULike[4], Mendeley[5], or Bibsonomy[6]. As a result, social bookmarking systems were taken into consideration as the third external source.

The goal of this dissertation was to explore ways to improve the recommendation by exploiting the information about target users collected from several kinds of external sources, and their combinations. In order to fulfill this goal, five studies and one external validity were conducted with two talk recommendation systems, CoMeT[7] and Conference Navigator 3[8], as study platforms. CoMeT and Conference Navigator 3 were developed to promote talks and to help users seek interesting talks. CoMeT is intended to promote awareness of interdisciplinary research talks between departments and between universities in Pittsburgh, mainly the University of Pittsburgh and Carnegie Mellon University. Conference Navigator 3 was developed as a community-based conference planner system, which aims to help conference participants go through the papers and add the most interesting papers to their schedule.

1.1 ORGANIZATION OF THE THESIS

This dissertation is organized as follows. In the following subsections of the Introduction, the research objectives are addressed, and then the issues and challenges of using external

[4] http://www.citeulike.org/
[5] http://www.mendeley.com/
[6] http://www.bibsonomy.org/
[7] http://pittcomet.info/
[8] http://halley.exp.sis.pitt.edu/cn3/

sources to improve research talk recommendations in small communities are discussed. Brief information about CoMeT and Conference Navigator 3 is presented and the external sources and the research objectives are explained. The second chapter reviews and describes related work on recommendation and using external sources for recommendations. The third chapter addresses preliminary analysis. The fourth chapter describes the proposed design space of recommendation approaches. The fifth chapter describes research questions, proposed research design, and research plans. The sixth chapter provides the results of study 1: external-source-augmented recommendation improvement. The seventh chapter provides the results of study 2: the cold-start problem. The eighth chapter provides the results of study 3: fusion recommendation. The ninth chapter explains the external validity results of the previous three studies. The tenth chapter describes the user study of external-source-augmented recommendations on the CoMeT system. The eleventh chapter describes the user study of transferring user models from the SciNet system to the CoMeT system. The twelfth chapter reviews, analyzes, and discusses results across different studies in respect to examined design space parameters. Finally, the thirteenth chapter discusses the summary and implications of this dissertation.

1.2 RESEARCH OBJECTIVES

This research aimed to explore approaches to the use of external sources in order to improve the quality of research talk recommendations. To accomplish this objective, a range of approaches using the external sources were developed and evaluated. Five studies, including three offline

cross-validations and two online user studies, were conducted in order to assess those approaches.

1.3 ISSUES AND CHALLENGES

To understand the challenges of using external sources to improve research talks, this section reviews some of the obstacles of research talk recommendations and the use of external sources.

Short-life Time Span Talks: Given the relatively short-life time span nature of research talks, even though some of them are well posted in advance, the system did not contain adequate information regarding user interests such as bookmarks or click logs. Talks associated with a "presentation date" have been considered special items with an "expiration date" (Minkov et al., 2010); the fact that after a presentation has passed it might be of little value being recommended to the users.

Under-contribution Communities: Although research talks were quite large in number, the number of bookmarks or ratings were too small. This is a common phenomenon of "under-contribution" social systems (Farzan et al., 2008). This resulted in a lower-than-expected quality of recommendations. Worse than that, many users rarely or never bookmarked talks. As a result, the system was not able to generate recommendations. Because data was sparse , the collaborative filtering algorithms could not produce an adequate amount of good quality recommendations.

Limited Content Analysis: Because of the short content length of research talks, the content-based recommendation finds it difficult to provide a sufficient recommendation by using just the content of research talks alone.

External Sources Crawling Limitation: Since social bookmarking scholarly papers services, such as CiteULike or Mendeley, impose the limitation on crawling, the crawlers cannot run throughout the whole content because there is a possibility that the dataset crawled might be missing some publications from the social bookmarking scholarly papers services.

1.4 RESEARCH QUESTIONS

The following are important research questions that need to be addressed:

1) Which recommendation approaches and which external sources can deliver the best improvement over the traditional within-system recommendations?
2) Could the external sources help alleviate the cold start situation in the research talk recommendations?
3) Which combinations of the different recommendation approaches generate better recommendation results?

In a realistic recommendation context, how do external-source-augmentation recommendations affect the relevancy and novelty of recommended talks?

1.5 COMET

A way of sharing information about research talks in colleges and universities is by posting paper flyers, announcing the talks on a dedicated department page, and sending e-mails to those people on mailing lists and to colleagues. While this approach may work well in a small college with well-positioned centers of expertise, it is not efficient in the context of large universities and, especially, proximity-located universities where talks on a similar subject could be organized by many different departments. This is especially the case in Pittsburgh where there are two large research universities, Carnegie Mellon and the University of Pittsburgh located within walking distance of each other.

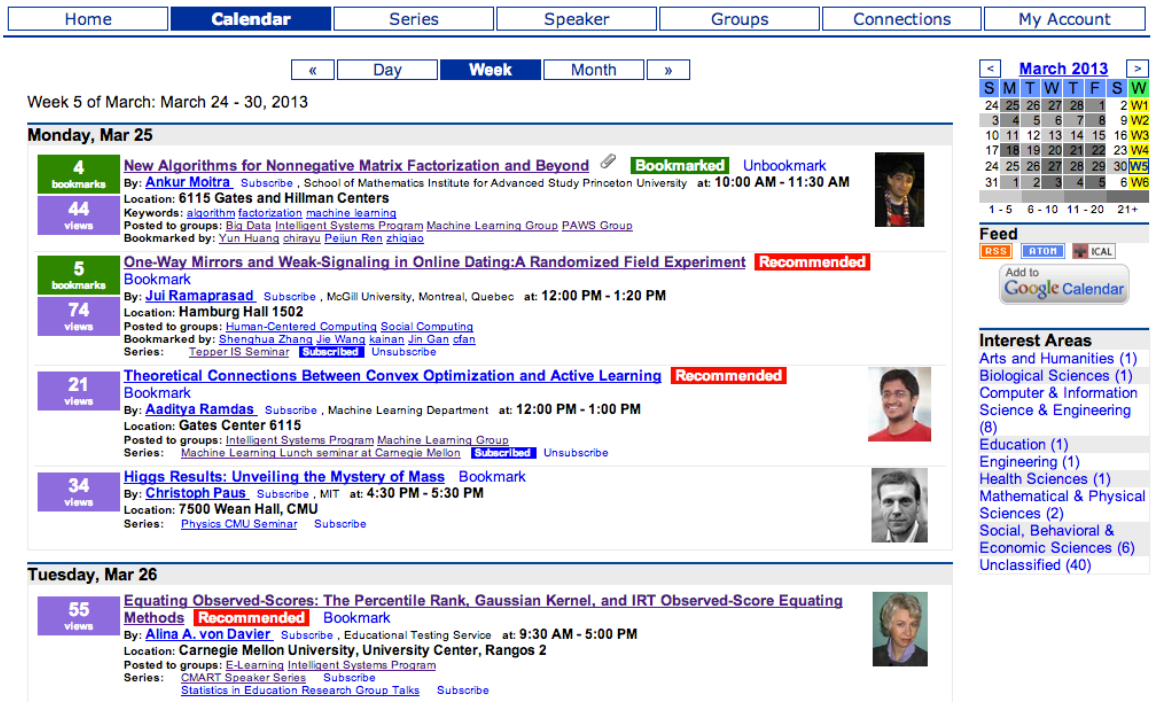


Figure 1: Screenshot of CoMeT Calendar

To improve the awareness of local researchers about relevant talks organized on both campuses, CoMeT was developed as a collaborative system for sharing information about research talks. The system was launched in the fall of 2009 as a typical collaborating tagging system, which allows any individual to announce, find, bookmark, and tag talks. CoMeT is a social system for sharing information about research talks in Carnegie Mellon and the University of Pittsburgh campuses. It supports both passive and active dissemination. Every user can post a talk by filling in a simple form. Mandatory fields include title, speaker, date, time, and location. The talk description field is not mandatory, but the talk abstract and speaker information is usually included in almost all posted talks. The users can browse talks from the calendar page (Figure 1) by date, standing series, organizing departments, and in other ways. The system can also disseminate talks using iCal, Google Calendar, Atom and RSS feeds. There are a variety of research talks hosted by CoMeT, ranging from physics, computer science, information science, economics, medical science, education, psychology, political science, and so on, mainly crawled from the websites of those departments.

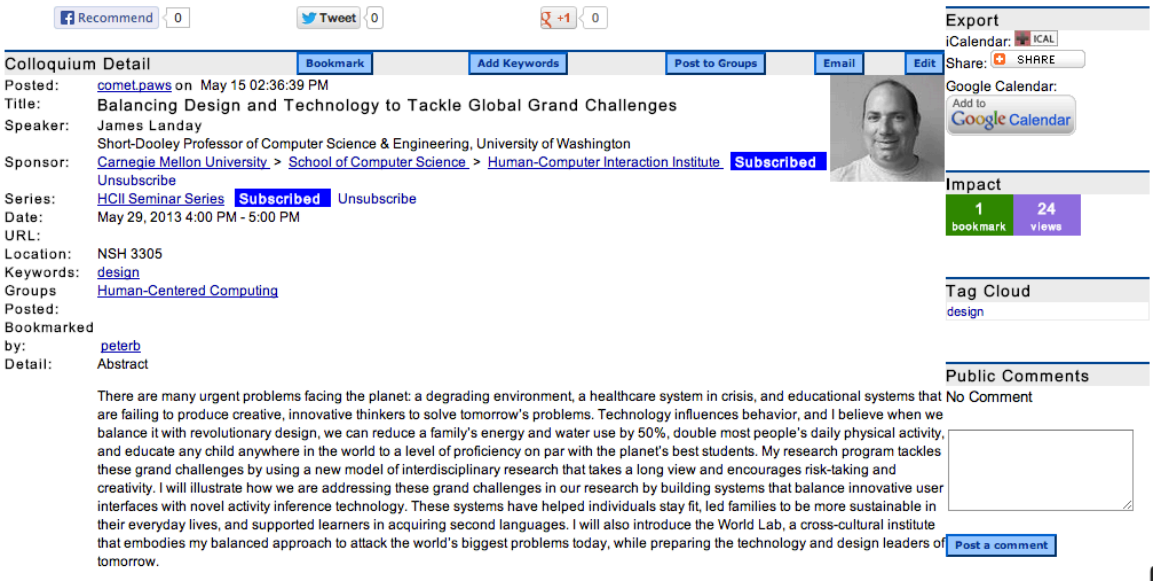


Figure 2: Screenshot of CoMeT Talk Detail Page

When users find interesting talks they can bookmark them, add tags, contribute to groups, and provide comments (Figure 2). The users can also share talks with their friends by email. The user cumulative activity related to a talk (viewing, tagging, sending by e-mail) is visualized in all lists where the talk is shown (Figure 1). This social link annotation feature is known as social navigation (Farzan, & Brusilovsky, 2008). It provides a simple way to use community wisdom for guiding users to good talks. The recommendation was implemented as a feature to help users locate talks that were of interest. The content-based recommender, which can build a profile of interest of individual users and recommend new talks immediately after talks are posted, is deployed. Recommendations in the current version of CoMeT are also displayed in the form of link annotation. For example, instead of showing all recommendations as a ranked list, the system adds a red tag, “Recommended”, to recommended talks in all contexts where a link to this talk is shown (Figure 1).

1.6 CONFERENCE NAVIGATOR 3

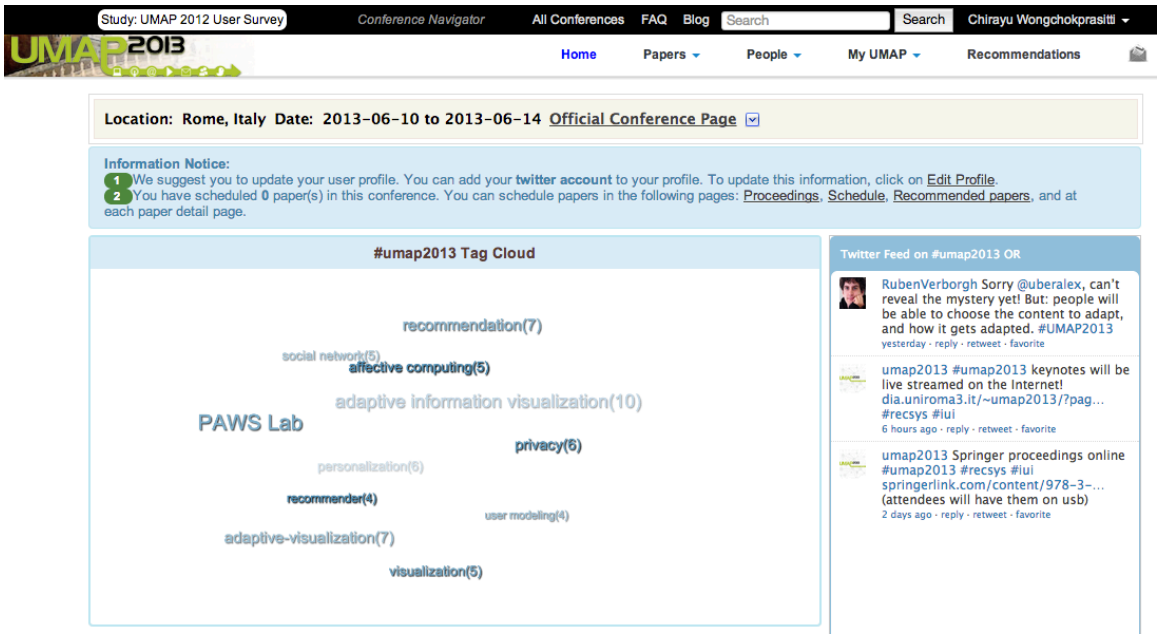


Figure 3: Screenshot of CN3 UMAP2013 Page

Conference Navigator 3 is an online social conference support system developed by the PAWS lab at the University of Pittsburgh. It allows conference attendees to explore the talks, schedule a personalized program of talks, or receive talk recommendations. Paper and pencil is a traditional way to plan which sessions to attend at an academic conference. Attendees do it by jumping between the conference program and proceedings, which provide more details about papers and posters. Large conferences are one of the venues suffering from this overload. Faced with several parallel sessions and large volumes of papers covering diverse areas of interest, conference participants often struggle to identify the most relevant sessions to attend.

The Conference Navigator 3 (CN3) system targets researchers attending a single academic conference by helping them to discover and share interesting papers, posters, and workshop papers presented during the conference. Conference Navigator 3 is the third generation in a series of developments on the shoulders of the Conference Navigator (Farzan & Brusilovsky, 2006) and the Conference Navigator 2.0 (Wongchokprasitti et al., 2010). It has been in operation since the summer of 2011, providing service for more than fourteen international conferences in five domains.

1.7 EXTERNAL SOURCES

In this dissertation context, external sources related to research interests are explained. The external sources contain the information that is available for retrieval by the user, including textual documents, images, music, audio, video, and software downloads as well as training, educational and reference materials. The content matching is used for the content-based recommendations (Ahn et al., 2007; Billsus & Pazzani, 2000; Sahebi et al., 2010; Wongchokprasitti & Brusilovsky, 2007) and, in addition, the content is used as content-boosted collaborative filtering (Melville et al., 2002). The co-occurrence of the same bookmarks in the scientific literatures from the social bookmarking services, such as CiteULike or Mendeley, is also used in the collaborative filtering. The external sources used in this thesis were divided mainly into four sources: personal webpage, user publications or bibliography, and bookmarked scholarly papers.

1.7.1 Personal Webpage

The personal webpage refers to World Wide Web pages constructed by individuals. The website content contains personal information about users and their interests. The personal webpage may not contain just one page but a collection of pages, separated by the user's judgment. This thesis exploited the personal webpage as a whole collection. The textual information was only held in the study only.

1.7.2 User Publications or Bibliography

User publications are scientific literature publications authored or co-authored by the user. The research used the title and abstract of the academic publications retrieved from Google Scholar and the Scopus scientific literature database. Falagas et al. (2008) stated that Scopus covered wider journals, but availability of recent articles (published after 1995) was limited compared to Web of Science[9]. The Google Citation service provides a means for users to confirm the authority of publications. If both are available, Google Citation is the first choice of study.

1.7.3 Bookmarked Scholarly Papers

There are online social bookmarking services, such as CiteULike, Mendeley, and Bibsonomy that provide the ability for users to bookmark, organize, and share scholarly papers. These sources provide external knowledge about their scientific interests. The title and abstract of the

[9] http://wokinfo.com/

bookmarked scholarly papers in the external sources, such as CiteULike or Mendeley, were exploited in the study. If both are available, the most recently updated source is preferred.

1.7.4 SciNet User Profiles and Search Logs

SciNet is an exploratory search system for scientific articles. It was introduced in 2013 (Ruotsalo et al., 2013). SciNet has indexed over 50 million scientific documents from Thomson Reuters, ACM, IEEE, and Springer. The exploratory search system in SciNet allows users to interact with an open user model. The open user model visualizes for interaction by organizing intents onto a radial layout where relevant keywords are close to the center of the visualization. After typing a query and receiving a list of documents, SciNet allows users to bookmark and then update the search by interacting with the user model. In each search iteration, the user model is inferred from the whole set of user actions; documents are searched based on the updated model, and the radial visualization is updated. The SciNet user model provides four sets of user search interactions, consisting of: (a) displayed scientific documents, (b) bookmarked scientific documents, (c) visualized keywords in the radial layout, and (d) manipulated keywords.

1.8 DEFINITIONS OF TERMS

Bookmarks: In this research, bookmarks are pointers that users utilize to indicate talks that are of interest. This research uses this knowledge as one of the signs of user interest to construct a user model.

Browsing Log History: Browsing log history refers to the list of pages a user has visited, associated with data such as page title and time of visit. The CoMeT system logs the user activities whether or not users log into the system.

CiteULike: CiteULike is a social bookmarking service for storing, organizing and sharing academic papers. The system allows users to add their academic references to their online profiles.

Co-Author Networks: Velden et al. (2010) define co-author networks as cooperative networks between groups of individual researchers publishing in specific research communities or domains.

Google Scholar: According to Wikipedia[10], Google Scholar is "*a freely accessible web search engine that indexes the full text of scholarly literature across an array of publishing formats and disciplines*". Google Scholar provides scholarly authors with the ability to update their personal information and publications in the Citation page, which other services such Microsoft Academic Search, Scirus[11], and CiteSeerX[12] do not allow authors to do.

Groups/Clusters: Groups are a set of users who share the similar topics or interests. When becoming a member of a group, this membership shows user interests or relevancy to the corresponding topic of the group.

Mendeley: Mendeley is a researcher social network for storing, organizing, and sharing academic papers and collaborating online with other researchers.

---

[10] http://en.wikipedia.org/wiki/Google_Scholar
[11] http://www.scirus.com/
[12] http://citeseerx.ist.psu.edu/

Metadata: The most common definition of metadata is data about data. In this context, metadata is structured data used to describe the research talk. In this particular paper, metadata can be viewed as a subset of data of the talk host, organization sponsors that support the talk, and the series to which the talk belongs.

Ratings: Rating is a scale classification assigned by users on certain items regarding the interest or attitude of users. Ratings have been the most popular source of knowledge for *Recommender Systems* to represent the preferences of users. In the CoMeT and CN3 systems, the rating scale is in the range from 1 to 5.

Recommender Systems: Recommender systems are systems that aim to help users deal with information overload by finding relevant items in a vast space of resources. They seek to estimate the 'rating' or 'preference' that a user would give to an item or social entity they had not yet considered.

Scopus: Scopus is a bibliographic database containing abstracts and citations for academic journal articles. Falagas et al. (2008) stated Scopus covered wider journals, but availability of recent articles (published after 1995) was limited compared with Web of Science.

Speakers: The speaker is the person who performs a presentation, such as a talk or paper. Occasionally, there are talks associated with more than one speaker.

Social Tags/Social Annotations: Social tags are free keywords attached to items that users share or items that are already available in the Social Tagging systems (STS).

2.0 RELATED WORK

This chapter reviews several research topics that have direct connection to the thesis: the recommender systems, cold-start problem, user profile representation, clustering approach, use of external sources to improve recommendation, and recommendation fusion.

2.1 RECOMMENDER SYSTEMS

Recommender Systems aim to help users deal with information overload by finding relevant items in a vast space of resources. Recommender systems have also been applied to a variety of different domains, such as music (Shardanand & Maes, 1995), Usernet articles (Resnick et al., 1994), email (Goldberg et al., 1992), and social bookmarking (Bogers, 2009). Goldberg et al., (1992) introduced the first recommender system, named Tapestry, almost 20 years ago. The system designed to take care of the increasing amount of emails user received. Resnick et al. (1994) and Shardanand and Maes (1995) introduced a new technique called Collaborative Filtering to provide recommendations based on previous actions performed by users and by like-minded others, denoted as nearest neighbors. Bogers (2009) made use of social tags to recommend bookmarks in the social bookmarking systems. Adomavicius et al. (2005) classified recommender systems into three main categories: content-based filtering, collaborative filtering,

and hybrid recommendation approaches. The content-based recommendation systems generate recommendations based on items that users prefered in the past. Collaborative filtering systems recommend items that people with similar preferences liked in the past. Hybrid systems are the combination of both approaches. Bogers (2009) and Parra and Sahebi (2013) classified recommender systems into four main categories: content-based filtering, collaborative filtering, rule-based or knowledge-based recommendation, and hybrid recommendation approaches. The rule-based or knowledge-based recommendation techniques allow the algorithm to reason about the relationship between a user and the available items.

2.1.1 Content-based Filtering

The root of the content-based approach to recommendation began in information retrieval (Baeza-Yates & Ribeiro-Neto, 1999) and information filtering (Belkin & Croft, 1992) research. Usually, content-based filtering for recommendation is approached as either an information retrieval or machine learning problem. In information retrieval, document representations have to be matched to user representations on textual similarity. In a machine learning problem, the textual content of the representations are incorporated as feature vectors, which are used to train a prediction algorithm. Many content-based filtering systems focus on recommending items that contain textual information. The textual content is used for construction of user profile representation that contain information about users' tastes and preferences. The profiling information can be obtained explicitly or implicitly. Explicit user information can be elicited by directly asking users or through questionnaires, for example. On the other hand, implicit feedback is information that was learned from their transactional behavior over time. One good

thing about content-based filtering algorithms is they do not require domain knowledge. Another is that content-based filtering algorithms are better at finding topically similar items than CF algorithms because they explicitly focus on textual similarity. Unlike collaborative filtering, the content-based recommendation uses the assumption that items having objective similarity will have the same ratings (Schafer et al., 2007). With this assumption, using the linear regression technique, the content-based rating prediction with user profiles between user u and item i is formulated as:

$$PredictedRating_{Content-based}(u,i) = \bar{r}_u + \frac{1}{n}\sum_{j=1}^{n}(r_{u,j} - \bar{r}_u) \times sim(i,j).$$

where $\bar{r}_u$ is the averaged rating regarding to user u, $r_{u,j}$ is the rating of item j regarding to user u, n is a number of total rated items from user u, and $sim(i,j)$ is the similarity between item i and item j.

2.1.2 User Profiles in Content-based Recommendation

The content-based recommendation mechinisms utilize the user profile representation as a way to make recommendations. NewsMe (Wongchokprasitti & Brusilovsky, 2007) is a hybrid 2-phase user-model unigram-based recommendation: one for the recent user interest (short-term), another for the user general preference (long term). The models are designed to be capable of representing a user's multiple interests in different periods of time, short-term interest and long-term interest. Even though researchers will not change their topic interests quickly, recent observations should be emphasized more than ones from long ago. Differentiation between short-term and long-term profiles has several desirable qualities in domains with temporal

characteristics (Chiu & Webb, 1998). The two-separated profile has been implemented in many applications such as adaptive new access (Billsus & Pazzani, 2000; Wongchokprasitti & Brusilovsky, 2007; Ahn et al., 2007) and personalized web search (Xue et al., 2009). Both user models use the title and detail as a bag of words and convert them into Term Frequency (TF) vectors then use the cosine similarity method to measure the similarity of two vectors. The predicted score is computed by averaging the weighted similarity of a new document with the most recent news stored in the user model. More detail concerning the user profile presentation is in section 2.3.

2.1.3 Collaborative Filtering

Collaborative Filtering (CF) is a process that filters and evaluates items through the opinions of the other people. The root of collaborative filtering is to make use of human nature by sharing opinions with others. Collaborative filtering recommender systems try to predict the utility of items for a particular user based on items previously rated by other like-minded users. For example, in order to recommend a movie to user *c*, the collaborative recommender system tries to find a group of users who share the similar movie tastes or preferences with user *c*. Then, user *c* gets only the movies that are most liked by this group of users.

According to Breese et al. (1998), collaborative recommendations algorithms can be grouped into two general classes: memory-based and model-based.

2.1.3.1 Memory-Based Collaborative Filtering

Memory-based Collaborative Filtering (Resnick et al., 1994) is a heuristics that makes rating predictions based on the entire collection of previously rated items by the users. The similarity measure between users is essentially a distance measure and is used as a weight. The similarity is a heuristic measure that is introduced in order to be able to find a set of "nearest neighbors" for each user. This approach has adavantages, such as recommended items are explainable to users, which is one important aspect of recommendation systems. It does not need to discern the content of the items being recommended. One of the common drawbacks of this approach is that its performance relies on the density of user item matrices. The quality of memory-based collaborative filtering recommendations decreases when the data is sparse. This mechanism can be categorized into two types based on similarity between users or items.

1. User-Based Collaborative Filtering

User-based collaborative filtering utilizes user ratings in order to compute similarity between users. GroupLens (Resnick et al., 1994) was the first system to implement user-based collaborative filtering recommendation. The user-based similarity is used for producing recommendations. The Pearson correlation (Lee Rodgers & Nicewander, 1988) and cosine similarity are the most well-known among many other similarities for this memory-based mechnism (Lemire & Maclachlan, 2005; Schafer et al., 2007). The equation of Pearson correlation *r* between user *u* and user *w* is quantified as:

$$r = \frac{\sum (u - \bar{u})(w - \bar{w})}{\sqrt{\sum (u - \bar{u})^2 \sum (w - \bar{w})^2}}.$$

While the cosine similarity measures the cosine of the angle between the vector space of user u and the vector space of user w as:

$$\cos(\vec{u},\vec{w}) = \frac{\vec{u}\cdot\vec{w}}{\|\vec{u}\|^2 \times \|\vec{w}\|^2} = \frac{\sum u \times w}{\sqrt{\sum u^2}\cdot\sqrt{\sum w^2}}.$$

This user-based collaborative filtering algorithm identifies the *k* most similar users to the target user by using the similarity mentioned earlier. Those similar users are used as a source in order to predict the ratings of an item in regard to those user-item matrices. A set of highest predicted rating items is recommended to the target user.

With variances of user ratings, one of the user-based collaborative filtering algorithms uses the rating average to adjust for users' mean ratings with the equation below:

$$PredictedRating_{User-based-CF}(u,i) = \bar{r}_u + \frac{\sum sim(u,w)\cdot(r_{w,i} - \bar{r}_w)}{\sum sim(u,w)}.$$

where $\bar{r}_u$, $\bar{r}_w sim(u,w)$, and $r_{w,i}$ are the average rating of user u, the average rating of user w, the similarity between user u and user w, and rating of the item i in regard to user w, respectively.

2. Item-Based Collaborative Filtering

The item-based collaborative filtering utilizes user ratings in order to compute similarity between items (Sarwar et al., 2001). Unlike the user-based collaborative filtering algorithm, the recommendation for any given item is based on a target user's ratings from similar items. The adjusted cosine similarity, the most popular similarity matric (Schafer et al., 2007), between item i and item j is formulated as:

$$\cos(\vec{i},\vec{j}) = \frac{\vec{i}\cdot\vec{j}}{\left\|\vec{i}\right\|^2 \times \left\|\vec{j}\right\|^2} = \frac{\sum i \times j}{\sqrt{\sum i^2}\cdot\sqrt{\sum j^2}}.$$

The advantage of the item-based machanisms over the user-based ones is scalability. The user-based collaborative filtering algorithms have a performance problem when the size of user item data is huge (Sarwar et al., 2001). While the user-based collaborative filtering algorithms exploit the nearest neighbors of users with respect to the target users to produce recommendations, the item-based collaborative filtering mechanisms use the ratings from all users whose rating of item i was provided. The rating item-based prediction can be formulated as:

$$PredictedRating_{Item-based-CF}(u,i) = \frac{\sum sim(i,j)\cdot r_{u,i}}{\sum sim(i,j)}.$$

where $sim(i,j)$ and $r_{u,i}$ are the item similarity between item i,j and the rating of item i in regard to user u.

2.1.3.2 Model-Based Collaborative Filtering

Unlike memory-based methods, model-based algorithms (Billsus & Pazzani, 1998; Hofman, 2003; Pavlov & Pennock, 2002) use the collection of ratings to learn a model. Billsus and Pazzani (1998) used the Singular Value Decomposition (SVD) technique as a way to reduce the dimensionality of the user-item rating matrix and find the interaction between users and items. Earlier studies (Golub & Reinsch, 1970; Klema & Laub, 1980; Van Loan, 1976; Wall et al., 2003) suggested that SVD is a method for transforming correlated variables into a set of uncorrelated ones to better expose the various relationships better among the original data items. SVD is a method for identifying and ordering the dimensions along which data points exhibit the

most variation. SVD takes a highly dimensional, highly variable set of data points and reduces it to a lower dimensional space, exposing the substructure of the original data more clearly, and ordering it from most variation to the least. Once it has been identified where the most variation is, it is possible to find the best approximation of the original data points using fewer dimensions. Recently, advanced techniques in collaborative filtering algorithms such as the matrix factorization model (Koren et al., 2009), provide better rating prediction performance by incorporating with the interaction between users and items. This model is closely related to the singular value decomposition (SVD) technique. More details are discussed in Koren (2008), Koren et al. (2009), Mnih and Salakhutdinov (2007), and Rennie and Srebro (2005).

Other techniques such as maximum entropy or probabilistic latent semantic analysis are exploited in model-based collaborative filtering. Pavlov and Pennock (2002) exploited the maximum entropy algorithm in order to attack the high data sparseness problem in the user-item matrix. Hofman (2003) used the Gaussian probabilistic latent semantic analysis as a framework to estimate the rating distribution and used the algorithm to predict the user ratings.

2.1.4 Knowledge-Based Recommendation

These approaches use rules and patterns, and they recommend items based on manually or automatically extracted knowledge of how a specific item meets a certain user profile (Burke, 2002). However, this rules and patterns restrict a recommender system from generating useless recommendations to users. For example, in the supermarket scenario, suggesting ones to buy milk is strange because the correlation of buying milk with any other item is so high. As a result, milk is an item that is always recommended to users. With this knowledge, the system

will not generate such useless recommendation (Burke, 1999). One of advantages of rule-based recommendation is that it does not suffer from the cold start problem. On the other hand, these approaches are limited in knowledge acquisition from users explicitly used to generate user profiles.

2.1.5 Hybrid Recommendation

There has been some prior research on integrating multiple recommendation methods, integrating collaborative filtering and content-based methods and using ontologies in recommendation systems as hybrid recommendation systems (Burke, 1999; Burke, 2002; Chung, 2007). Combining collaborative filtering and content-based approaches is one way to avoid limitations of content-based and collaborative systems. There are many examples of memory-based hybrid (Billsus & Pazzani, 2000; Melville et al., 2002; Pazzani, 1999) approaches to combine collaborative and content-based techniques. Model-based hybrid approaches incorporate one component as a part of the other (Basu et al., 1998) or build one unifying model (Schein et al. 2002; Wang & Blei, 2011). Interestingly, Wang and Blei (2011) used the Latent Dirichlet Allocation topic-model approach to demonstrate a hybrid approach to make recommendations. More detail about different ways to combine approaches is discussed in section 2.6 recommendation fusion.

2.1.6 Recommendation in Academia Domain

The focus of this thesis is recommendation of *research talks*, which has not been explored well in the past. However, some works has been done in the related areas of recommending research papers. This section briefly reviews both areas of recommendation. Earlier studies (Bollacker et al., 1999; Bollacker et al., 2000; Gori & Pucci, 2006; McNee et al., 2002; Minkov et al., 2010; Sahebi et al., 2010; Wang & Blei, 2011) on recommendation of scholarly papers have been conducted for a decade. Bollacker et al. (1999; 2000) introduced a system for tracking scientific literature that is relevant to the research interest of users as part of CiteSeer[13] digital library. They used unigram-matching, co-occurrences of citation links, and metadata associated with scientific literatures. McNee et al. (2002) exploited the collaborative filtering technique in order to recommend research papers using citation web between papers for creating the ratings matrix. They investigated four different collaborative filtering algorithms for the citations selection, which consisted of co-citation matching, user-item, item-item, and Naïve Bayesian Classifier. The user-item and item-item collaborative filtering algorithms performed very well in the offline experiments, but did not do well in the online experiments because in the online experiment, they did not allow algorithms from recommending citations if the authors were the same as users. The best algorithms from this study provided either very relevant or very novel recommendations, although there was no single algorithm that provided both at the same time. The follow-up study (McNee et al., 2006) showed the user-user collaborative filtering and content-based term-matching TF-IDF algorithms outperformed the Naïve Bayesian and Probabilistic Latent

[13] http://citeseer.ist.psu.edu/

Semantic Indexing (PLSI) classifiers but the results also suggested that the qualities of algorithms regarding to users varied according to task. Gori and Pucci (2006) introduced the random walk based scoring "*PaperRank*" algorithm by exploiting citation graph, resembling the PageRank algorithm (Page et al., 1999).

Brusilovsky et al. (2010) and Sahebi et al. (2010) studied the effects of augmentation of information into research talk recommendations in academic conferences and open lectures at University of Pittsburgh and Carnegie Mellon University by using additional information of users from the CiteULike social bookmarking system or social tags users assigned to research talks. The evaluation result showed improvement in the quality of recommendations. Minkov et al. (2010) also studied future events recommendation in academia. Their study compared a content-based filtering using *RankSVM* (Joachims, 2006) with a proposed collaborative filtering *LowRank* algorithm. Two user studies were conducted at Massachusetts Institute of Technology and Carnegie Mellon University over 15 weeks and concluded that *LowRank* had a better performance. Wang and Blei (2011) used the Latent Dirichlet Allocation topic-model approach to demonstrate a hybrid approach as a combination of content-based and collaborative filtering to make recommendations.

2.1.7 Cold-Start Problem in Recommender Systems

Cold-Start (Pavlov & Pennock, 2002; Schein et al., 2002) issues are classic problems for recommender systems. The conditions are that new items are introduced or there are new users or users whom little or no information has yet been acquired. For those conditions, the recommender systems encounter a difficulty generating a good quality of recommendation.

There are two types of cold-start problems: new-system and new-user (Middleton et al., 2002). The new-system situation is when no user has provided ratings yet, and therefore, there are no user profiles. The new-user situation is when the system already has a set of user profiles and ratings, but no available information regarding to a new user. One limitation of content-based methods is the new user problem. The users have to provide enough information for a content-based recommender system to really understand the user's preferences and present them with reliable recommendations. Consequently, a new user, having very little information in the system, would not be able to get accurate recommendations. Collaborative filtering approaches face the same problem as content-based ones. The collaborative systems need the sufficient number of ratings the new users provide in order to generate a reliable recommendation. The next problem for collaborative filtering is the new item problem. When the new item added to the system, collaborative filtering cannot generate any recommendations on new ones because collaborative systems rely only on the previous user preferences and ratings. Lastly, data rating sparsity is another classic problem. The ability of prediction of ratings from a small number of examples is important but a critical mass of users to make it happen is another factor. With the growth of the number of various social and personalized systems that maintain user profiles, the idea of using profiles from one system to improve the quality of personalization in another system is becoming more and more popular (Mehta, 2008; Sahebi & Brusilovsky, 2013).

2.2 USER PROFILE REPRESENTATION

A user profile is a collection of personal data associated with a specific person. There are three types of user profiles: unigram or keyword profiles, semantic network profiles, or concept profiles (Gauch et al., 2007). There are two ways of user profile construction (Schaffer et al., 2010): through direct explicit user information feedback (Ahn et al., 2007; Godoy et al., 2004; Wongchokprasitti & Brusilovsky, 2007; Zigoris & Zhang, 2006), or through implicit monitoring of user activities (Hong et al., 2009; Matthijs & Radlinski, 2011; Sugiyama et al., 2004).

Explicit feedback requires extra efforts from the users while monitoring user activities can raise privacy concerns. User profiles should be able to be updated or incremented dynamically and be able to represent the short-term profiles and the long-term profiles. The short-term profiles represent the current interests of the users and are subject to frequent update over time while the long-term profiles represent the general preferences of users (Gauch et al., 2007; Xue et al., 2009). Liu et al. (2002) and Xue et al. (2009) suggested that the level of granularity of profiling (such as the individual-level, the group-level, the global-level, and the hybrid) can be used to construct the user profile as well.

2.2.1 Unigram Profiles

Unigram-based profiles are the most common user profiles (Gauch et al., 2007). The unigram terms are extracted from the documents users visited (Bogers, 2009), web pages bookmarked (Matthijs & Radlinski, 2011; Sugiyama et al., 2004), or directly provided by the users (Ahn et al, 2007; Dasan, 1998). Weighting unigrams in regard to a certain function, such as frequency, can

be used to identify the level of importance of certain unigrams given resources such as visited web documents.

Dasan (1998) filed a patent titled “Personalized information retrieval using user-defined profile”. Dasan used the terms that were manually inputted and exploited them in order to create a single unigram user profile. The patent explained the newspaper generator with the user-defined profile is able to produce the personalized newspaper. The patent also provided the user interface for users in order to add/remove unigram into their user profiles. Chen and Kuo (2000) used the single unigram profile in order to make a personalized web search. The adaptive system will re-ranked the retrieval results based on comparing the semantic relevance and co-occurrence of searching terms with the unigram in the user profile. Matthijs and Radlinski (2011) used the user browsing history to create the single unigram profile. Their study explored the use of the attributes of content of browsing documents, such as title, keywords, and metadata, in order to construct different user profiles and re-rank them to generate the personalized web search.

The aforementioned simple unigram vector profiling (Chen & Kuo, 2000; Dasan, 1998) does not handle the dynamic of the user behaviors. As a result, the multiple unigram vectors profiling is adopted in the adaptive system recently. Multiple unigram profiling is a technique to create multiple unigram vectors per interest instead of just one vector in the one unigram vector profiling method. YourNews (Ahn et al., 2007) and NewsMe (Wongchokprasitti & Brusilovsky, 2007) are examples of the adaptive systems that provide the personalization on the news access service by exploiting the multiple unigram vector user profile. In particular, YourNews uses the open user model, allowing users to control the user model. Its open user model allows users to add/remove unigrams in each news topic. The user profile in the YourNews system has two profiles for each news topic: short-term and long-term. NewsMe (Wongchokprasitti &

Brusilovsky, 2007), a personalized news access system, also represents the user profiles by comprising the short-term positive and negative profiles and the long-term profile on each news topic. Users explicitly provide ratings on viewed news content. NewsMe also uses the decay function in order to give more weight to current feedback documents than old ones in the long-term profile. The TF-IDF unigram weighting in the unigram profile is the most common scheme in order to weight terms with more frequeny in the document (Term Frequency) but discount them if they appear too often in the corpus (Inverse Document Frequency). The TF-IDF can be calculated by equation below:

$$TF - IDF(t_k, d_j) = TF(t_k, d_j) \cdot IDF(t_k, d_j).$$

$$TF(t_k, d_j) = \frac{f_{k,j}}{\max_z f_{k,j}}$$

$$IDF(t_k, d_j) = \log \frac{N}{n_k}$$

where t_k stands for any term or keyword, $f_{k,j}$ is the frequency of term t_k in the document, d_j. $\max_z f_{k,j}$ is the maximum frequency of any term in the document d_j. N is a number of documents in the corpus and n_k is a number of documents that contain the term t_k. To compute the similarity between the user profile u and the document d being recommended, the most common way is to use the cosine similarity to measure it as follows:

$$\cos(u, d) = \frac{\sum u \times d}{\sqrt{\sum u^2} \cdot \sqrt{\sum d^2}}.$$

2.2.2 Semantic Network Profiles

Collecting positive and/or negative user feedback is a standard way of building semantic network profiles (Gauch et al., 2007). The semantic network profiles represent a weighted network in which each node represents a concept. Semantic web profiling is used to extract the keywords from user-feedback documents. The difference with the keyword profiling is extracted keywords are added to a network of nodes instead of vectors. Nodes represent individual words. In more complicated approaches, nodes represent a concept and its associated words.

Gentili et al. (2003) introduced the online digital libraries document filtering system named InfoWeb. InfoWeb uses semantic network user profiles representing long-term user interests. The user profiles are a semantic network of concept profiles. Users can provide explicit feedback back to the system. In the initial state, each semantic network contains unlinked nodes. Each node represents a concept, called planet. Planet contains one weighted term representing that concept. In order to create stereotypes that represent the scope of the digital library, they cluster documents in the digital library into *k* categories. To ensure that the clustered categories are semantically meaningful, domain experts choose *k* documents as seeds for the initial state of the clustering algorithm. The new users are asked to give positive/negative feedbacks about the stereotypes. The top weighted keywords are extracted from the rated documents. The system creates a single semantic network representation for the new user. WIFS (Micarelli & Sciarrone, 2004), a personalized web search, builds on a single semantic web representing interests of the user. The web search results are filtered from the AltaVista[14] search engine. User profiles

[14] http://www.altavista.com/

comprise three parts: a header, a set of stereotypes, and a list of interests. A header includes the personal data of a user. A stereotype, or a stereotypical user, consists of a set of interests. A frame of slots represents a stereotype. Each slot stores three facets: domain, topic, and weight. The semantic links include lists of co-occurrence keywords in the slot-associated document. Co-occurrence keywords are associated to a degree of affinity with the topic. The user profile is seen as a set of semantic networks. A slot is a planet and semantic links are satellites. The WIFS uses the specific-domain experts to identify a set of terms and store them in databases as the most relevant for each topic of interest. Recently, Fossati et al. (2012) introduced the news recommendation based on the semantic network profile. Their system is intended to make the entertainment-domain news recommendation to the generic users or stereotypical users. The TMZ[15] news articles are used to create a semantic network.

2.2.3 Ontology/Concept Profiles

Ontology or concept user profiles are similar to semantic web user profiles in that both are represented by a network of conceptual nodes (Daoud et al., 2010; Liu et al., 2002; Middleton et al., 2001). However, the concept-based user profiles consider nodes representing abstract topics not in the specific words or in a set of related words. Also, the concept user profiles are constructed similarly with unigram user profiles as they are represented as vectors of weighted features. In concept profiles, features represent concepts rather than words or a set of words. The simplest ontology user profiles are constructed from a reference taxonomy or thesaurus. Most systems are based on a reference ontology or taxonomy. A subset of concepts and relationships

[15] http://www.tmz.com/

are extracted from use profiles. Concept profiles are based on subsets of existing concept hierarchies because creating such board and deep ontology profiles are expensive and they are mostly manually built. The concept profiles provide richer representations. They also allow better interest tracking and propagation.

QuickStep (Middleton et al., 2001) is an example of a system that employs the concept user profile. QuickStep is a research paper recommender system that uses the ontology user profile based on mapping the user browsing behaviors on research papers with the hierarchical topic ontology from Open Directory Project (DMOZ)[16] taxonomy of computer science topic. The simple decay function is also deployed in order to keep the user profile focused on the current user interest. Liu et al. (2002) built a personalized web search that constructs a user profile and general profile by mapping the user search history with the concept hierarchy category from the Open Directory Project. The general profile representation is the global profile of users from the search history of the system. These two profiles are used to generate the personalization of the web search. Daoud et al. (2010) exploited an ontology profile generated from the Open Directory Project. They introduced the personalized graph-based document-ranking model with the concept user profile. The re-ranking method is used to personalize the search results. The cosine similarity measurement is used to calculate between the retrieval document and the top weighted concepts of the user profile.

[16] http://dmoz.org/

2.3 CLUSTERING APPROACHES

In the granularity of user profile construction, the higher level beyond the individual user model is the group model. The group model or cluster model groups users who share similar interests. This subsection provides a literature review about the clustering approach. The clustering approach is an unsupervised classification of observations, feature vectors, and/or patterns. The goal of clustering is "to separate a finite unlabeled data set from a 'natural'; defined as hidden data of unobserved samples generated from the same probability distribution", into a finite and discrete set (Xu & Wunsch, 2005). Indeed, the clustering plays a significant role in data analysis in many research domains, underlining the important issue of how we make use and improve the clustering method. Unfortunately, there is no universal clustering theory widely accepted, yet (Milgan & Hirtle, 2003). Broadly, clustering can be categorized into two approaches: heuristic-based clustering and model-based clustering approach.

2.3.1 Heuristic-Based Clustering Approach

Heuristic-based clustering approach is an experience-based method to optimize a clustering problem by using iterative processes to refine a candidate clustering solution. For example, k-means technique changes the centroid position in the clusters after finishing the approximation process. This approach has advantages in no requirements or a few assumptions relate to the clustering problem. The disadvantage is whether or not there is any existing cluster, the approach would cluster the observed data. However, the approach does not guarantee an optimal solution. There are two subdomains of the heuristic-based clustering approaches.

2.3.1.1 Hierarchical Clustering

A hierarchical clustering is widely used in social network, biology, engineering, and marketing because of its ability to construct a taxonomy chart (Fortunato, 2010; Jain & Dubes, 1988). The hierarchical technique is a technique for transforming the proximity matrix into tree-like nested partitions. Its idea is based on the objects and the nearby objects rather than faraway objects. The clusters are determined by the distance measurement. There are two major types of hierarchical algorithms: the agglomerative and the divisive. The agglomerative clusters are iteratively merged if their proximity is close. And the divisive clusters are split iteratively if their similarity is low. The common representation of a hierarchical clustering is a tree structure, called a dendrogram as depicted in Figure 4. The dendrogram enables a data analyst to see how objects are being merged into clusters or split at successive level of proximity. The data analyst can decide that the entire dendrogram describes the data or can perform a tree cutting line of the dendrogram at an appropriate level of proximity (Kettenring, 2006) such as the elbow technique (Thorndike, 1953).

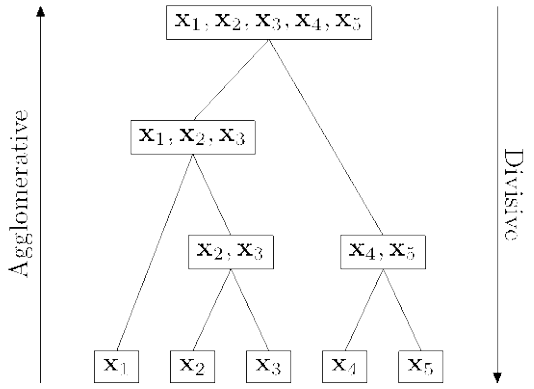


Figure 4: Dendrogram (Gan et al. 2007, Figure 7.1)

2.3.1.2 Partitional Clustering

Partitional clustering is an appropriate method in the representation of a large dataset (Jain & Dubes, 1988). The principle concept of partial clustering is to separate the data objects in the clusters by optimizing a given cost function (which is defined by a distance between points, and/or points to centroids). Specifically, each cluster contains at least one object and each object belongs to only one cluster. This requirement could be relaxed when considering the fuzzy (overlapping) clustering (Fortunato, 2010). The widest use of partitional algorithm in the literature is *k-means* clustering (MacQueen, 1967). The given cost function in the *k-means* clustering is defined by the intra-cluster distance or squared error function. It starts from an initial distribution of *k* centriods, which can be randomly assigned. Each object data is assigned to the nearest centroid regarding the cost function. Then, the new centroids are recalculated and so on. With a certain number of iterations, the positions of centroids are stable. The *k-means* technique provides intuition of how clustering works and it is easy to implement. The result also can be improved by performing *k-means* clustering with different starting distributions of k centroids, and the solution is determined by the clustering that produced the minimum value of the cost function.

2.3.2 Model-Based Clustering Approach

The model-based clustering approach offers a principal way to hthe euristic-based approach. The underlining assumption of this approach is that every cluster data distribution is in the form of a certain distribution. The model-based clustering approach assumes a mixture of probability distributions of data, and each data could be represented in more than one clustering

(Ketternring, 2006). The model-based clustering framework provides an efficient way to deal with several problems in this approach such as a number of clusters (component densities), initial values of parameters, and distributions of clusters (e.g., Gaussian distribution).

The approach needs a model for clustering and methods to quantify the approach. The approach uses certain models for clusters and tries to optimize the fit between the data into the models (Fortunato, 2010; Gan et al., 2007). Constraints and geometric properties of the covariance matrices are used to measure the quality of the clustering. Among model-based clustering approaches, the expectation-maximization algorithm (EM) is the most widely used with model-based clustering approach for estimating the parameters of a finite mixture probability density. The model-based clustering framework provides a way to deal with several problems in this approach such as a number of clusters (component densities), initial values of parameters, and distributions of clusters (e.g., Gaussian distribution). Other approaches such as Self-Organizing Map (Kiang, 2001; Kohonen, 1999; Smith & Ng, 2003; Vesanto, & Alhoniemi, 2000) clustering approach exploit the neutral network approach in order to cluster the data and determine a number of clusters.

In Hancock et al. (2007), the latent position cluster model captured the social networks' three key characteristics: transitivity, homophily on attributes, and clustering.

- Transitivity feature is when two actors have ties to the third actor; they are more likely to be tied than not.
- Actors who have similar attributes such as age, gender, geography, race, etc. are more likely to be tied than others who do not. They call this tendency "homophily on attributes."
- By the nature of social networks, they show clustering characteristic whether they are driven by transitivity, homophily on unobserved attributes, or the position in the network.

In their experimental results of the second method, Bayesian model using Markov chain Monte Carlo (MCMC) sampling, which estimates latent position of actors in social space and the cluster model at the same time, showed the better results in clustering in both the American monks monastery dataset, and the adolescent health dataset than the two-stage maximum likelihood model. The first method estimated the latent position then computed the cluster model.

2.3.3 Clustering Approach Selection

In Chapter 14, Gan et al. (2007) stated that the problems in the heuristic-based approach were selecting a "good" clustering method, and deciding the "correct" number of clusters. On the other hand, the problems in the model-based clustering approach were reduced to the model selection problem in the probability framework. The factors leading researchers to select one or another approach are as follows:

Heuristic-Based Approach

1) The number of clusters must be pre-defined and not difficult to choose.

2) The population distribution of observed data cannot be determined.

3) The heuristic-based approach can handle special data types such as spatial datasets, multimedia, etc. It can perform efficiently on the arbitrary-shape clusters and spatial structures in the density-based clustering algorithms.

4) The heuristic-based approach is fast.

Model-Based Approach

1) The population distribution of observed data can be determined.

2) The model-based approach cannot handle special data types and can only perform on the convex-shape of clusters.

3) Hierarchical clustering in the initial iteration state requires storage and time that has faster growth rate than linear relative to the size of the initial partition. As a result, it cannot be directly applied to large data sets (Fraley and Raftery, 1998).

Number of Clusters Selection

The criterions such as Bayesian Information Criterion (BIC), Akaike information criterion (AIC), and other penalized likelihood criteria, are use to calculate the number of clusters (component densities). The BIC penalizes the complexity of the model where complexity refers to the number of parameters in the model (Handcock et al., 2007, Sec. 4).

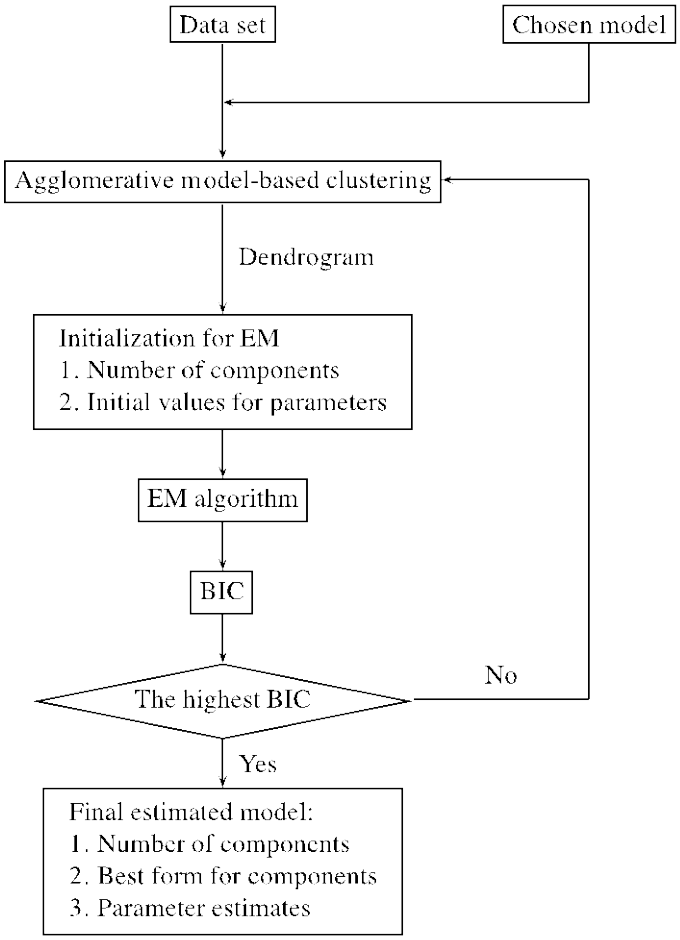


Figure 5: The Flowchart of the Model-Based Clustering Procedure (Gan et al., 2007, Figure 14.1)

2.4 USE OF EXTERNAL SOURCES TO IMPROVE RECOMMENDATION

There has been prior research on using external sources to boost recommender systems. In the area of recommender systems, this multi-system approach and integration of user models built by other personalization systems is called user model mediation. User model mediation (Abel et al., 2013; Berkovsky et al., 2008; Carmagnola et al., 2011) has been studied in recent years.

Berkovsky et al. (2008) used a user model mediator to integrate user models, which imports user models collected by personalized systems. Sosnovsky et al. (2009) provided multiple ways for integrating user models in adaptive learning systems. Cross-item mediation is a popular technique applied to existing recommendation techniques, such as content-based filtering (White, 2002), collaborative filtering (Rafter & Smyth, 2001), and hybrid approaches (Chung et al., 2007). This mediation technique makes recommendation based on the items similar to the past items, including cross-items from the other remote systems users liked. There is an increased amount of information about the users due to the opportunity to benefit from the efforts provided by other systems (Heckmann et al., 2005). This leads to increased coverage in terms of qualitative and quantitative improvement of the user model since the aggregated user model can cover more aspects, including the features that one system could not acquire by itself. This improvement has the potential to give better adaptation results. Semeraro et al. (2009) used a dictionary, Wikipedia[17], and user-generated content in order to create a set of user knowledge and exploit it to boost the content-based recommendations. Katz et al., (2011) used the Wikipedia webpages as an external source to improve recommendation accuracy in the collaborative filtering approach by mapping its pages with items being recommended and quantifying their similarity. Their approach was to match the items to the corresponding pages in Wikipedia and use other the pages' metadata to help compute the item-to-item similarity. Katz et al. (2011) claimed that their approach worked even when data sparsity was high.

[17] http://www.wikipedia.org/

2.5 RECOMMENDATION FUSION

Past research indicated that fusing several sources of information in making recommendations could increase the quality of recommendations. This section reviews the work in this direction.

Table 1: Description on Recommendation Fusion

Fusion Method	**Description**
Cascade	One recommender refines the recommendations given by others.
Feature Augmentation	One method is chosen to compute a set of features, which is used as parts of inputs for another method.
Feature Combination	Features are extracted from different knowledge sources and combined together. One method uses these features to calculate recommendation.
Meta-level	A meta-level hybrid model is a model learned by one approach and used as input to another approach.
Mixed	Recommenders present their recommendations together.
Switching	The recommender system switches the methods, depending on the situation.
Weighted	The scores of recommendation methods are weightily combined together.

The approach to combining different recommendation algorithms has been studied in the last decades in the machine learning, data mining, and information retrieval domains. One of the main reasons is that each algorithm has its own strong characteristic (Burke, 2002). Fusion, combing differing strengths, has been a technique utilized in search of better performance. Burke (2002) identified seven strategies to combine different recommendation approaches.

While these seven combination methods have their advantages and drawbacks, there has not been a systematic comparison of them (Bogers, 2009). As a result, it is difficult to make a conclusion about which technique is most effective in which situation. Most of the related work on recommendation fusion has focused on combining content-based filtering with collaborative filtering (Adomavicius & Tuzhilin, 2005; Basu et al., 1998; Billsus & Pazzani, 2000; Pazzani, 1999).

In machine learning, the combination and hybridization methods are similar but with different names from those classified by Burke (2002). The first is *bagging*. Bagging is similar to Burke's weighted combination method. The bagging method is to take a vote or weighted vote in order to generate such combined recommendation. Breiman coined the term "bagging" in 1996. Breiman explored the properties of bagging for classification and numeric prediction in theory in empirical ways. A second one is randomization. The idea of randomization is to generate diversity among ensemble of classifiers by introducing randomness from the input into learning algorithms. The methods, such as random forests (Breiman, 2001) and rotation forests (Rodriguez et al., 2006), are considered to be the randomization method. Third, *boosting* involves combining models that complement one another. The boosting method learns an optimally weighted combination of a set of weak classifiers to produce a strong classifier (Freund & Schapire, 1996). Another technique is *stacking*. Wolpert (1992) introduced stacking

or stacked generalization, where the output of one classifier is used as an input feature for the next classifier. It corresponds to Burke's feature augmentation method.

In order to combine the output of different recommendation algorithms, there are two different things that need to be taken into account: the scores or ratings of each recommended item, or the ranks of the items in the list. These two options are commonly referred to as *score-based fusion* (Lee, 1997; Montague & Aslam, 2001) and *rank-based fusion* (Liu et al., 2007) in the related work. Studies (Lee, 1997; Renda & Straccia, 2003) reported there was no agreement to which method was better. In this study, both data fusion methods are being explored.

2.5.1 Score-Based Fusion

In score-based data fusion, different algorithms yield a different kind of similarity value. To compare with a different range of values, scores need to be transformed to be comparable one another (Croft, 2002), usually by normalizing them. The score normalization method (Lee, 1997; Montague & Aslam, 2001) is needed to apply for each result to map the score into range [0,1] as in the equation below:

$$score_{norm} = \frac{score_{original} - score_{\min}}{score_{\max} - score_{\min}}.$$

where $score_{norm}$, $score_{original}$, $score_{min}$, and $score_{max}$ stand for the normalized score, the original recommendation score, the minimum score, and the maximum score in the list of recommendations, respectively.

2.5.2 Rank-Based Fusion

Similar to the score-based fusion, the rank-based fusion also needs to be normalized the rank of recommendation list as equation below (Renda & Straccia, 2003):

$$rank_{norm} = 1 - \frac{rank_{original} - 1}{|rank|}.$$

where $rank_{norm}$, $rank_{original}$, and |rank| are the normalized rank, the original rank from the recommendation list, and a number of recommendation items in the list. According to Lee (1997), *CombSUM* and *CombMNZ* produced the best result in score-based data fusion.

In Information Retrieval, this area has been discussed under the name of "data fusion". There are two main approaches to results fusion: to combine retrieval results generated using *different query representations* but with the same algorithm, and to combine retrieval results generated using the same query, but with *different algorithms*.

2.5.3 Fusion on Different Representations with the Same Algorithm

The first method of data fusion is by combining retrieval results generated using *different query representations* but with the same algorithm (Belkin et al., 1995; Ingwersen & Järvelin, 2005). Belkin et al. (1995) showed that combining different boolean query formulations improve retrieval effectiveness. Ingwersen and Järvelin (2005) fused different query and document representations from a cognitive IR perspective. The principle of *polyrepresentation* was presented as each query, searcher, and retrieval model can be seen as a different representation of the same retrieval process.

Xue et al. (2009) introduced the collaborative personalized search by using two-step data fusion. The first step was to use a global user model as a way to smooth the unseen terms in the individual models. Then, the group models were incorporated with the combined model from the first step. The two-step fusion is depicted in Figure **6**.

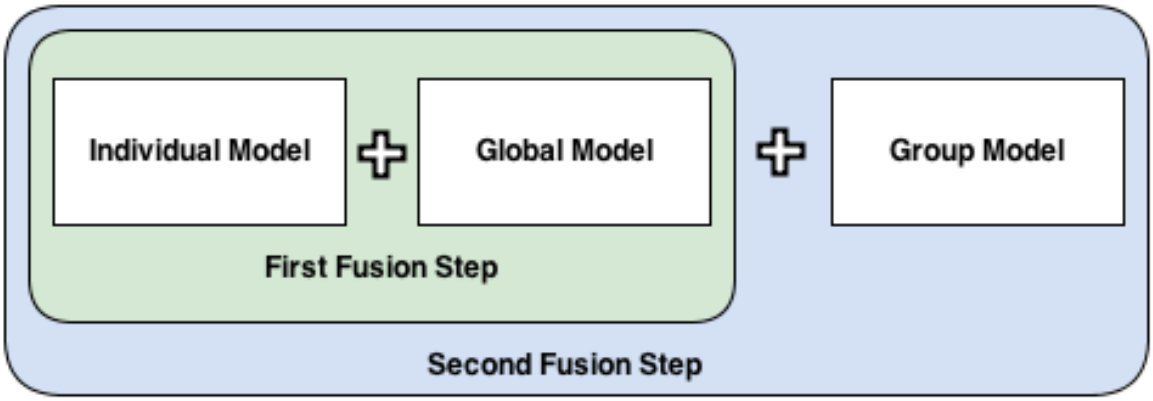


Figure 6: Two-step Data Fusion

Also, Xue et al. (2009) found out that when data is very sparse, using global models as smoothing was more accurate than the group models.

2.5.4 Fusion of the Same Query but with Different Algorithms

The second one is by combining retrieval results generated using the same query, but with *different algorithms* (Choochaiwattana, 2008; Croft & Thompson, 1987; Lee, 1997). Croft and Thompson (1987) combined a vector space model with a probabilistic retrieval model. Lee (1997) examined retrieval runs examined in different settings and found that *CombSUM* and *CombMNZ* were the best-performing methods. The *CombSUM* method combines retrieval runs

by taking the sum of similarity values for each document separately, and the *CombMNZ* boosts the sum by the number of runs that actually retrieved the document. Choochaiwattan (2008) used social tags from Delicious[18] corpus to augment indexing and exploited social tags in many setting to re-rank the results with no significance compared with the results from Google.

Interestingly, Chiu et al. (1998) studied the utility of including two separate models in the context of student modeling. Their research showed that data representation and learning algorithms differ significantly from the text classification approach. They assumed that currently collected data would reflect the recent knowledge or preferences of users more accurately than data from previous time periods. They also found that they were not significantly different when combining more than two models.

2.6 CROSS-SYSTEM RECOMMENDATION

In recent years, cross-system and cross-domain recommendation research has grown in popularity due to the growing number of personalized systems that collect information about the users. In this context, it becomes natural to export information collected about a user in one system and to transfer it to another system (maybe in a different domain) to improve the quality of recommendations. It has been argued that this transfer might be most valuable in cold-start context when a recommender system has insufficient information about new users in a target system (Sahebi et al., 2013). Despite the overall interest in this field, there are still very few studies exploring real cross-system user information transfer due to the lack of sizable datasets

[18] http://delicious.com/

that have pairs of users in two systems. As a result, a major share of research on cross-domain recommendation focused on transfer learning approaches that do not assume a common user set for different domains (Li et al., 2009; Pan et al., 2011).

According to the most recent review (Fernadez-Tobis et al., 2012), the work on cross-domain recommendation could be split into two large groups: those using collaborative filtering and content-based approaches. The classic examples of collaborative filtering transfer are Berkovsky et al., 2008 and Cremonesi et al., 2011. These authors offered an extensive discussion of cross-domain recommendation problems and suggested interesting generic approaches, but they were not able to explore these approaches in a true cross-domain context, instead using instead artificial datasets produced by separation of single-domain user movie ratings into subdomains. More recent work explored collaborative transfer approaches in more realistic settings with hundreds of users having ratings in both domains (Sahebi et al., 2013). Content-based cross-domain recommendation appeared to be a harder challenge. No immediate success was reported for simple keyword-based profile transfer approaches. As a result, the majority of research and success in this category focused on using shared semantic-level features such as social tags (Shi et al., 2011) or Wikipedia (Loizou, 2009). Unfortunately, it leaves open the case where user preference data in the source domain includes no tags and can't be associated with an extensive ontology such as Wikipedia.

3.0 PRELIMINARY ANALYSIS

This research explored the use of external sources of information to improve research talk recommendation. The primary challenge was how to benefit from the use of the external sources. To explore the feasibility of using external sources and their prospective value, a preliminary user study was conducted. The study explored the value of using one specific external source, user bookmarks from CiteULike system. The study also explored different approaches to user profiling. CoMeT system was chosen as a system framework for the study. Eight users of CoMeT system who also have CiteULike accounts were recruited. All users were PhD students at the School of Information Sciences, University of Pittsburgh. The recommending approaches using external sources were explored in the preliminary study as in the following section.

3.1 RECOMMENDATION APPROACHES

In this study, two approaches were explored to improve recommendations in CoMeT:

1) By using the title and abstract of bookmarked papers from CiteULike in addition to the standard use of the title and abstract about bookmarked talks from CoMeT, and
2) By using tags for better representing information about talks (and user interest) in addition to standard use of text-only information from talk descriptions. In addition, the

fusion approaches, for example, using both kinds of information (descriptions and tags) from both systems were explored.

While fusing information from two systems is relatively straightforward (simply combining them as a bag of words), combining tags and text in both item representations and user profiles is not obvious and can be done in several ways. The next section introduces several user profile representations, which were explored in various ways to fuse unigram terms and tags. Then, the recommendation approaches based on these representations are discussed.

3.1.1 User Profile Representations

There are many ways to combine various sources of information for building user profiles. Tags and unigram terms in abstracts and titles of CoMeT talks and bookmarked papers from CiteULike have been utilized. To construct user profiles, the following representations are presented:

Keywords Only (KO): To represent talks in this model, only keywords extracted from talks' titles and abstracts are used. Each talk is considered as a bag of words and represented as a vector in unigrams vector space weighted by TF-IDF weighting.

All bookmarked papers from CiteULike can be represented as a $k \times l$ matrix D_c (k is the number of papers from CiteULike and l is the number of unigram terms used in those papers) and each CoMeT talk is represented in an $e \times m$ matrix D_t (with e as total number of talks and m as total number of unigram terms). To integrate these two sources of data in this model, a

$(k + e) \times (l + m - o)$ matrix D is obtained, where o is the number of mutual unigram terms between two CoMeT and CiteULike systems.

Keywords+n*Tags (KnT): In this model, tags are considered as regular unigram terms and each talk is treated as a bag of words containing the talk's abstract, title, and tags. Each tag appears n times in this bag of words. Each talk is represented as a vector in unigrams and tags vector space weighted by TF-IDF weighting scheme. Merged talks and papers are performed using the previous model.

Keywords Concatenated by Tags (KCT): In this model, tags are considered a separate source of information and are treated separately. A bag of unigram terms and a bag of tags are distinctly obtained for each talk and paper. Using TF-IDF weighting scheme, a tag vector and a unigram vector is built for each talk. Next, each talk is represented by concatenating unigram and tag vectors as one vector in unigram and tag vector space ($d_c = (w_{1,c}, w_{2,c}, \ldots, w_{l,c}, t_{1,c}, t_{2,c}, \ldots, t_{j,c})$, where $w_{i,c}$ shows the weight of i^{th} unigram terms in document c, l is the total number of unigrams, $t_{i,c}$ shows the weight of i^{th} tag in document c, and j is the total number of tags).

In this case, each CoMeT talk is represented as an $e \times (l + j)$ matrix (D_t), where e is the number of CoMeT talks, l is the total number of unigrams in CoMeT, and j is the total number of tags in it. Also, CiteULike paper is represented as a $k \times (m + i)$ matrix (D_c), where k is the number of CiteULike talks, m is the total number of unigrams in CiteULike, and i is the total number of tags in it. After merging these two matrixes, a matrix D has its $(e + k) \times (m + I + l + j - o - p)$ dimension, showing all talks and papers in unigram terms and tag vector space, where o and p are respectively the number of common unigram terms and tags between two systems.

To study the impact of various external sources on recommendation systems, each of the aforesaid models are utilized once by:

1) Using only CoMeT data,
2) Using both CoMeT and CiteULike data.

3.1.2 Recommending Talks to Users

In order to recommend research talks to users, the *K*-nearest neighbor method is exploited. In this method, top *K* closest talks to the user profile are recommended to each user. User profiles are built based on users' bookmarked and rated talks and papers. Each user's bookmarked and rated talks and papers, is weighted by user ratings, in a vector in talks and papers vector space. To obtain unigram-based user profiles, representations are constructed as mentioned in the previous section. Unigram-based user profile (*UP*) is obtained by multiplying the vector of user talks in talk vector space (*U*) by the matrix of talk unigram terms represented in unigram and tag vector space (*D*):

$$UP = U.D'$$

The generated user profile (*UP*) is a vector consisting of user's related unigram terms (and tags), weighted based on the importance of each unigram (or tag). To measure the distance between talks and user profiles, the cosine distance measure is in use.

3.2 EXPERIMENTAL SETTING

The recommended talks returned from recommending approaches were merged, randomly ordered, and presented to the user for evaluation. To evaluate each item in the merged recommended list, a user had to answer three questions measuring relevance to research interest, overall interest, and novelty:

1. Is this talk related to your interest? (yes/no question)
2. How interesting this talk to you? (5-point scale)
3. If the talk is related to your interests, how novel is this talk to you? (5-step scale)

Altogether, there were five models to compare: KO, KnT (with n = 1, 2, 5), and KCT. Each model was used twice, once to recommend talks using only CoMeT data and one using both, CoMeT and CiteULike.

Table 2: Preliminary Metrics

Performance Aspect	Metrics
Relevance	Precision @ position k
Overall Interest	nDCG
Novelty	Average Rating @ position k

Three measures (to explore relevancy, interest, and novelty correspondingly) were computed for every model for each position in the top-10 recommendation list as shown in Table

2. Non-relevant recommendations were considered as having zero novelty. To reduce the volume of reported data, in all tables, results are only shown for the best *n* value for KnT model, which was n=1.

3.3 EXPERIMENTAL RESULTS

3.3.1 Relevance

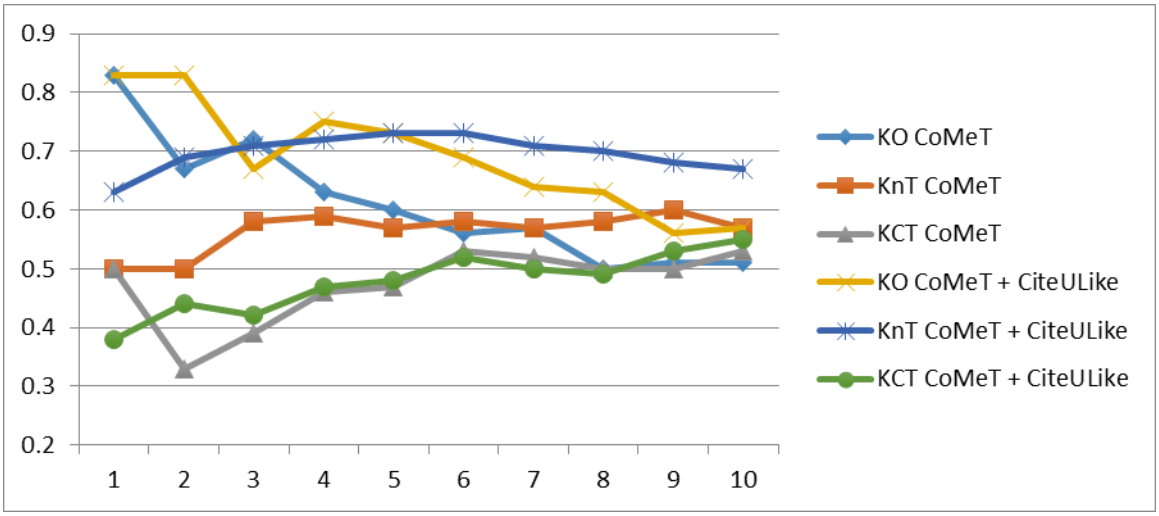


Figure 7: Precision Results for Different Models with Different Number of Recommendations

As illustrated in Figure 7, adding tags using the fusion approach (KnT), resulted in a better cumulative precision for the top 10 recommendations. The exceptions were the very top positions in the recommendation list where KO model worked better in both cases. The different behavior of these two models caused this interesting effect. For KO model, the precision was

high, at the top positions, but then dropped rapidly. The precision of KnT model was more stable and overran KO at position 5.

Augmenting CiteULike data in both KnT and KO models apparently increased the precision. Both tags (with KnT and CiteULike papers) could help with the precision. Moreover, these two effects seemed to be stackable: KnT model which included both CoMeT and CiteULike data had the best cumulative precision.

At the same time, adding tags using KCT model degraded the system's precision. This might be because of high dimensionality of the vector space model when the unigram vector was concatenated with the tag vector. In this case, the distance between documents and user profile increased and decreased the variance between similarities of user profile to different talks. Different sources of information may result in less precise results if integrating them in the wrong way.

3.3.2 Interest

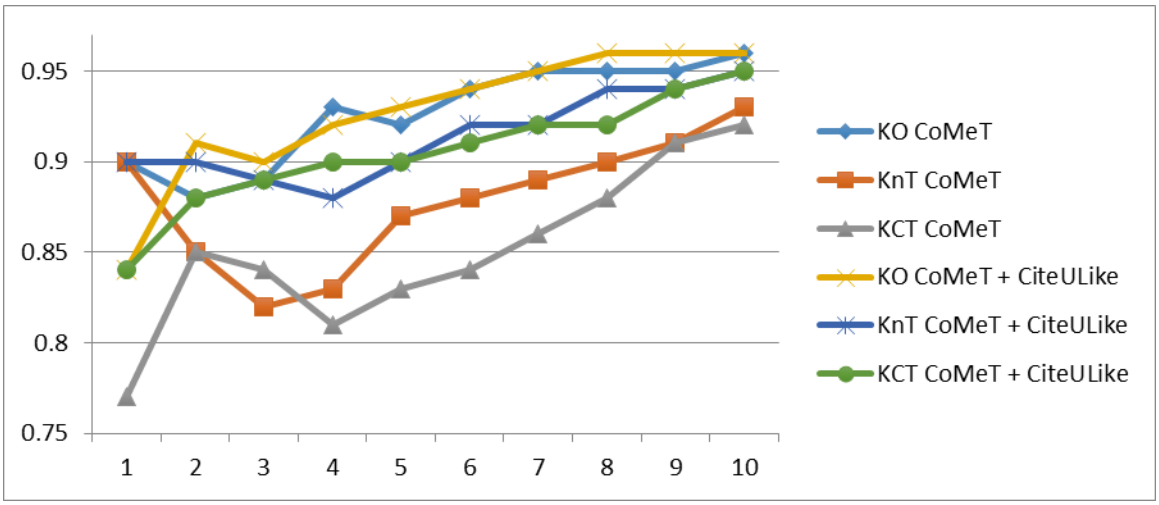


Figure 8: Overall Interest Results (nDCG) for Different Models with Different Number of Recommendations

Figure 8 shows that "overall interest" results are a bit different, compared to the previous measure. The positive effect of using CiteULike data was expected. In both, KCT and KnT models, augmenting CiteULike data into CoMeT data boosted the user cumulative interest in recommendations, although there was a very small difference between the KO model and KO + CiteULike model. Interestingly, the no-tag-in-addition KO model both with and without CiteULike data produced best results. These results were very close to each other and better than other models, in general.

To explain these results, the "relevancy" term, which was understood as a fit to the user research work was deliberately separated from the "overall interest" term, which was understood as an overall attraction of an item. From the CoMeT user observations, many users seemed to be interested in some talks on general topics (like art and politics), which had little in common with

their research interests. This is a separation from relevance and interest allowed cases, where a talk is rated as interesting, yet non-relevant. The analysis of questionnaire data confirmed that the observations were correct. There were a number of talks in their user profiles for almost all users. The decreasing ability of models to recommend interesting talks with the tags augmentation could be caused by the decreasing ability of models to recommend interesting, but not relevant talks. This was a natural outcome of user tagging behavior, which was focused mostly on their research interests. As a result, a simpler KO model was able to better grasp overall user interests.

3.3.3 Novelty

In **Figure 9**, the novelty evaluation of recommending approaches resulted in the opposite direction. Here, adding tags with KnT fusion models provided the largest positive impact. Both KnT models (with or without using CiteULike data) produce more novel recommendations to users. However, the impact of using CiteULike data was not consistent. The KnT model fusing with CiteULike data generated recommended talks more novel to users. This result shows that adding different sources of information, especially tags in this case, can improve the novelty of recommendations. Novelty results can be explained by a broader range of vocabulary of tags, provided by users, for describing talks or papers, compared to unigrams from their content. Each user uses tags to describe a talk or a paper from his/her own point of view, which might be different from the terms included in the document's abstract or title.

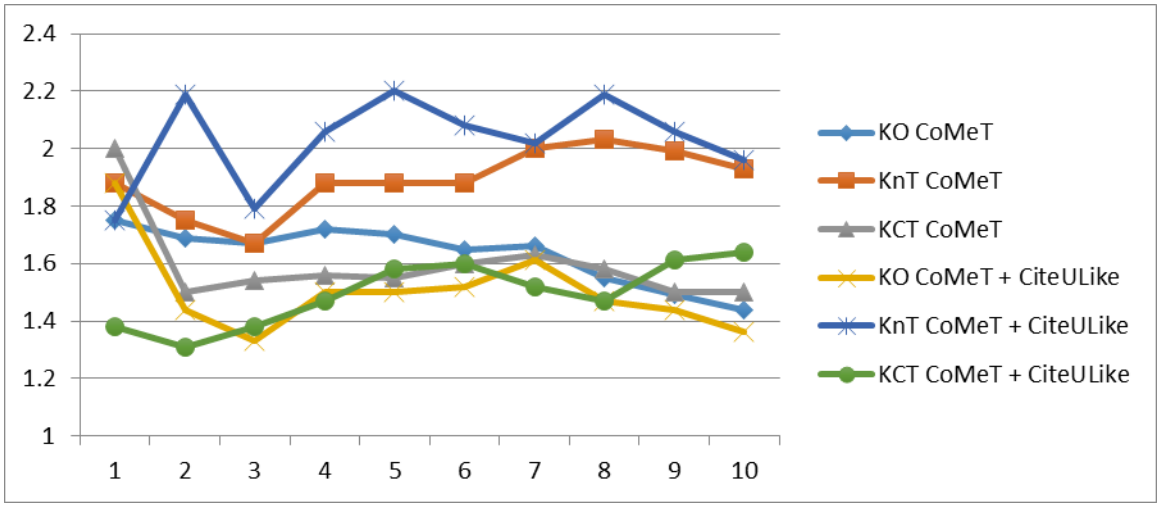


Figure 9: Novelty Results for Different Models with Different Number of Recommendations

On the other hand, augmenting-with-CiteULike-data KO model decreased the average of novelty of talks recommended to users. This is due to the distinctive natures of CoMeT and CiteULike systems. Users usually use CiteULike for adding, reviewing and rating related papers to their research field. This requires a user to spend a noticeable amount of time on a paper, and users prefer to review related papers to their field of research. On the other hand, CoMeT contains information about talks happening within a specific time given on a particular date. It is more plausible for a user to bookmark a more interesting, more novel, even less relevant talk, knowing that he/she might miss this amount of information given in a limited time. As a result, user profiles from the CoMeT system included a wider area of user interests, compared to user profiles from the CiteULike system, which usually contains more relevant documents.

3.4 DISCUSSION AND SUMMARY

The three evaluation results show several trends. First, the addition of tags in the KnT fusion model helped improve both novelty and relevance of recommended talks. This effect was more prominent for novelty. In contrast, when considering only the user overall interests, both no-tag-in-addition KO models produced slightly better results. As for both KCT concatenation models, they seemed to have real problems. The concatenation models decreased the quality of recommendations for all kind of measures, producing worst performances in most of the cases.

The effect of adding CiteULike was more consistent. It typically produced better results for all measures, although its effect for interest measure was negligible. Interestingly, the effects of adding tags and adding unigram terms appeared to be stackable. For example, an approach that uses both tags and CiteULike "stacks" the separate effects of the component approaches, resulting in the best approaches for relevance and for novelty measures.

One preliminary conclusion was that including another reliable user profile into the existing user profile would increase the precision of recommendations but the way to augment the additional profile is important. Preliminary results may be different for the separate injection features. Especially, when taking tags into account, the augmentation method should be more concerned. Otherwise, in some cases, the recommendation performance might be reduced.

Moreover, the relevancy of recommended talks, measured by the precision, increased for all the models that augment with CiteULike data, while the overall interestingness of recommended talks varied when models were augmented by including CiteULike data.

Adding tags to the models increased the novelty of recommendations generated from the models that use both CoMeT and CiteULike data. However, it increased their relatedness in larger

number of recommendations. As a preliminary conclusion, injection of unigram terms from another source of data, to obtain relevant content-based recommendation was more reliable than including tags into the models. On the other hand, including tags from various sources of information into the model yielded more interesting or novel recommendations.

4.0 THE DESIGN SPACE FOR RECOMMENDATION APPROACHES

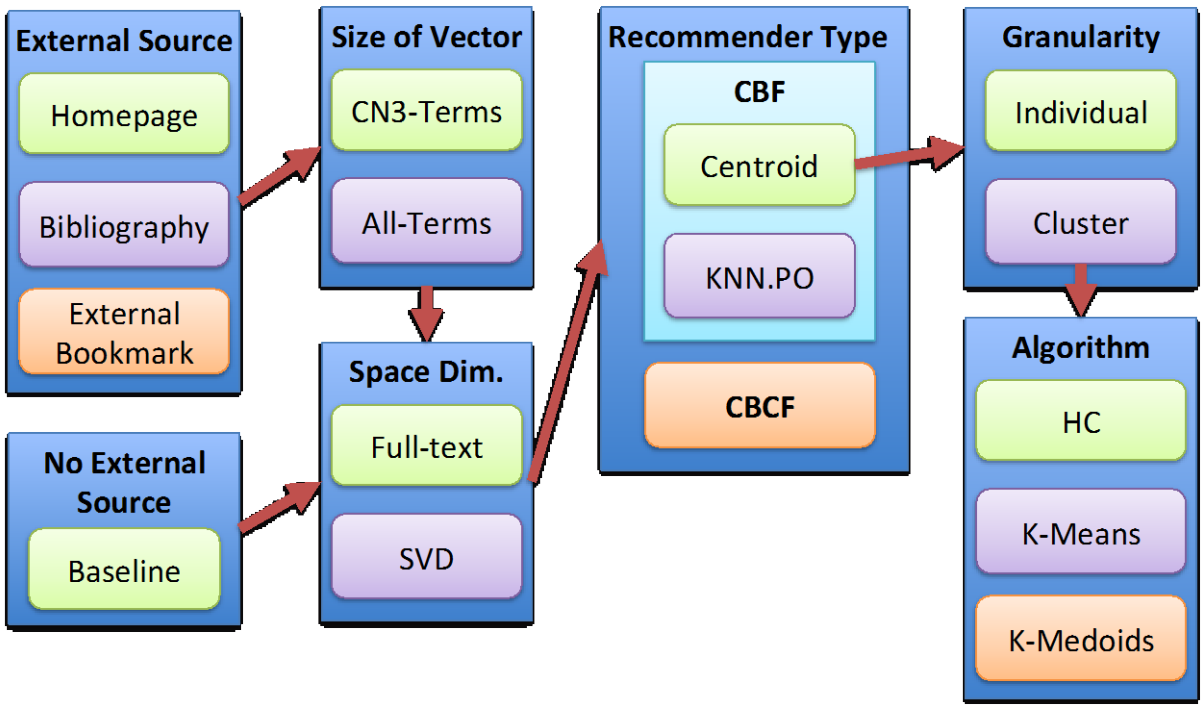


Figure 10: Exploring Design Space

This research explored the use of external sources for improving the quality of research talk recommendations. The recommender algorithms explored in the study differ by five key parameters: the source of augmentation (homepage, user publications, external bookmarks, or none), the profiling approach (unigram terms vs. SVD user profiling), profile granularity (the individual-level user model vs. group-level user model), points of user interests (centroid vs.

positive-sample-only *k*-Nearest-Neighbors) and the type of recommender approach (Content-based Filtering vs. Content-boosted Collaborative Filtering). The goal was to explore all meaningful combinations of these parameters and to find which worked best (See Figure **10** for the whole design space). In addition, hybrid approaches that fuse the results produced by basic approaches, combining the different recommending approaches were explored.

The first dimension of the study was to examine and compare the quality of recommendation performance after user profile augmentation with each type of external sources. The second dimension was to study the impact of different user profile representations from the unigram user profile to the latent semantic space user profile from the singular value decomposition (SVD) technique. The third dimension was to examine the impact of the user profile granularity on the recommendation quality by comparing individual profiles against group-level profiles compiled for like-minded group level. The fourth dimension was to study how the choice of specific recommender approaches would affect the recommendation. Finally, the fifth dimension was to study the impact of fusing in different recommendation approaches. These exploration options were discussed in details below.

4.1 TYPES OF EXTERNAL SOURCES

There are four types of external sources:

1. *Homepage*,
2. *User publication or bibliography*,
3. *External scholarly paper bookmarks*.
4. *SciNet user profiles and search logs*

One of the goals of the study was to explore how augmenting the user model with each of these sources would affect the quality of recommended talks. In other words, the quality of recommendation performance after the user profile augmentation by each type of the external source was compared. In addition, this research studied how the quality of recommendations on the cold-start situations would be impacted by augmentation of different external sources. Also, we would like to explore the effect of fusion of recommendations from different external source augmentations. Lastly, the effect of transferring the user model from SciNet to CoMeT was investigated.

4.2 USER PROFILING

4.2.1 User Profile Representation

In the study, two principally different user profiling approaches, classic unigram profiles and modern LSI-based profiles were compared. With the unigram user profile, the similarity between

user profile and the documents being recommended is relied solely on matching the exact same vocabularies. Even documents can discuss very similar topics but the similarity might be very low. In information retrieval, one way to solve this problem is to use the Latent Semantic Indexing (LSI) technique (Deerwester, 1990). The LSI is based on the dimensionality reduction via SVD. Billsus and Pazzani (1998) showed that the SVD technique could help boost the accuracy in the collaborative filtering. SVD is a well-known matrix decomposition technique that converts *l*x*n* matrix *B* into three matrices as follows:

$$B = U \times \Sigma \times V^T$$

where, *U* is an *l*x*l* matrix of *B* taken from the orthonormal eigenvectors of BB^T, Σ is an *l*x*n* rectangular diagonal matrix, *V* is an *n*x*n* matrix of *B* taken from the orthonormal eigenvectors of B^TB, and V^T is the transpose. With the purpose of finding the latent spaces of the content, the U matrix was the only one that was used. SVD not only helps reduce matrix dimensions, but also help the model find the latent semantic relationship in the dataset. The computational time is decreased as well.

The SVD vectors extracted from U matrix were exploited as the LSI-based (SVD) vector space document representation and used in the SVD user profile representation in the centroid CBF and CBCF models. The SVD vector space document representation is a vector certain *k* latent topics from the *U* matrix with the same row as the document located in the original *B* matrix. The SVD profile for each user is an arithmetic mean of certain *k* latent topics from the *U* matrix regarding documents users bookmarked or rated. The *k* number of latent topics starts from 100, stepping up 100 until the latent topics number is 2000. In the study, the first 2000 latent

spaces from *U* matrix were used as the user profile representations to explore the proper number of latent topics.

4.2.2 User Profile Granularity

In addition to using different kinds of user profiles, the usefulness of expanding the individual user profile into a group level profile in order to create a stereotypical model was investigated. One problem of using only individual profiles is that the model encounters data sparseness and cold-start. Since new users have no bookmarks or some users bookmark only a few of research talks, it is difficult to generate the recommendations with good quality. The cluster models increase the density of data, especially for users with few bookmarks. The group-level model alleviated problems while the individual model cannot handle those problems. Recently, the group-based approaches were used as a supplement to personalization (Xue et al., 2009). Inserting the external sources into the individual model can help decrease the data sparseness problem. However, it is interesting to see whether the group-level model can increase the recommendation performance, compared to the individual ones.

4.2.3 User Profile Application

In this dissertation, applications of user profile representation were investigated. The single-user-interest and multiple-user-interest user profiles were compared to each other. The centroid user profile approach was selected as a representative of the single-user-interest user profile representation. The centroid approach is an arithmetic mean position of vector space of all

documents in the user profile. The centroid vector is computed to represent one center point of user interests.

The *k*-Nearest-Neighbors (KNN) and positive-sample-only *k*-Nearest-Neighbors (KNN.PO) were chosen as applications of the multiple-user-interest profile representation. The KNN profiles contain positive samples (bookmarked talks) and negative samples (non-bookmarked talks), while KNN.PO stores only positive samples (bookmarked talks). The KNN and KNN.PO models in this dissertation recommend talks in two steps. First, for each talk in the test set, KNN or KNN.PO models search for its *k* nearest samples in the profiles. KNN profiles find samples regardless of sample type (positive or negative) while KNN.PO profiles look for only positive ones. Second, models calculate a cosine similarity between a target talk and its talks in the user profile, and its signs, depending on the sample type (positive for bookmarked talks or negative for non-bookmarked talks). Then, KNN and KNN.PO models average those similarities and assign it to a target talk as its weight. Finally, models rank all target talks based on their weights. In this dissertation, a number *k* was explored, starting from 5, stepping up 5 a time, until 100. The KNN models were exploited only in Chapter 11.0 : Study 5, while KNN.PO models were used in all CBF studies.

4.3 TYPES OF RECOMMENDER

This study explored two types of recommender approaches that could be used in our specific context: content-based filtering recommendation (CBF), and content-boosted collaborative filtering (CBCF). The characteristics of recommendations vary upon the nature of type of

recommender systems. CBF would generate recommendation more specific in the similar content. Novelty is the drawback of this recommender. In contrast, CBCF would provide more novelty of recommendations. This dimension of the study aimed to study whether augmentation with different external sources on the different type of recommender system performed differently.

4.4 FUSING DIFFERENT RECOMMENDATION APPROACHES

In the information retrieval domain, the collaborative personalized search using this fusion model (Sugiyama et al., 2004; Xue et al., 2009) has been applied and generated consistently better and more robustly than the individual model. The cluster user model combines the user profiles of like-minded users, making a group-level recommendation. Using the individual model in recommender systems faces the data sparseness problem and over fits model. Combining another model can help alleviate those problems (Xue et al., 2009).

In this thesis, the modified score-based data fusion methods from Lee (1997) were proposed. The modified *CombSUM* method is the combination of weighted score or weighted rank of recommendation on each item as in the equation below:

$$CombSUM(i) = \sum_{n=1}^{N} w_n \times score\_or\_rank(i, r_n)$$

where w_n, r_n, and $score\_or\_rank(i, r_n)$ are the weighted score on each algorithm, the r_n recommendation method, and the normalized score or rank of item i from the recommendation r_n, respectively. The modified *CombMNZ* method is the way to fuse by multiplying the sum of

weighted score or weighted rank of recommendation with a number of recommendation algorithms the item is recommended as below:

$$CombMNZ(i) = h(i,R)\sum_{n=1}^{N} w_n \times score\_or\_rank(i,r_n)$$

where $h(i,R)$, w_n, r_n, and $score\_or\_rank(i,r_n)$ are a number of recommendation methods that return item i, the weighted score on each algorithm, the r_n recommendation method, and the normalized score or rank of item i from the recommendation r_n, respectively.

There are many possible combinations for fusing recommending approaches but in order to keep the fusing study feasible and align with the main goal of this thesis, using external sources to improve research talks recommendation, two methods of data fusions were used in this research:

1) Fusion within the same source,
2) Fusion between the different sources.

The fusion between the different sources was chosen because it is more compelling to study different ways to fuse multiple sources instead of trying to make use of one source with different conditions. The detail of data fusion is discussed in Chapter 8.0 .

5.0 RESEARCH DESIGN

To answer the research questions, a series of five related experiments based on two different experimental designs were conducted. The overall goal of the experiments was to determine whether external sources could be used to improve the user experience by providing better ranked lists of recommended talks compared to those provided by baseline recommender systems.

The first three experiments applied the *experimental design 1*: offline cross-validation. The experimental design 1 was intended to use the five-fold cross-validation on Conference Navigator 3 (CN3) system platform. The last two experiments applied the *experimental design 2*: online user study. The experimental design 2 was planned for the two user studies on CoMeT system platform. In this section, the research questions, hypotheses, the experimental designs, and details of the experiments are described.

5.1 EVALUATION METRICS

To evaluate the performance of experimental models versus baseline approaches, this research used

1) Accuracy assessments made by the cross-validation evaluation on the proposed approaches against the baselines in the CN3 platform system,
2) Relevance, Attending, and Novelty measurement made by subjects, who were recruited in the user study, on the selections of the experimental approaches against the baseline in the CoMeT platform system.

The evaluations were the recommendation *Accuracy* on experiments 1 – 3, the recommendation *Relevance* and *Novelty* on experiment 4, and the recommendation *Attending* and *Novelty* on experiment 5. The Mean Average Precision (MAP) was used to measure the accuracy performance of the proposed models against the users in CN3 and CoMeT platform systems in all the experiments. In addition, the Normalized Discounted Cumulative Gain (NDCG) was calculated to measure relevance, attending, and novelty of experimental models against the users in CN3 platform system in studies 4 and 5, and to assess relevance, attending, and novelty of the experimental models against the subjects' ideal ratings in studies 4 and 5. In these experiments, the collective ratings of users from the CN3 platform system and subjects in the CoMeT platform system were considered ideal.

5.1.1 Mean Average Precision (MAP)

A classifier labels examples as either positive or negative in a binary decision problem. The confusion matrix or contingency table represents the decision made by the classifier. There are four categories in confusion matrix: true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN). True positives are examples correctly labeled as positives. False negatives are negative examples incorrectly labeled as positives. True negatives refer to negative

examples correctly labeled as negatives. Finally, false negatives correspond to positive examples incorrectly labeled as negatives.

Table 3: Confusion Matrix

	Actual positive	Actual negative
Predicted positive	*TP*	*FP*
Predicted negative	*FN*	*TN*

Given the confusion matrix, the *precision at rank k*, *prec(k)*, matric can be defined as equation below:

$$prec(k) = \frac{TP}{TP + FP}.$$

The precision at rank *k* is the number of correct items up to rank *k*, divided by *k*. The *non-interpolated average precision* of the ranking is the average of *prec(k)* for each position k_i as in the equation below:

$$Avg\,\Pr c(k) = \frac{1}{n}\sum_{i=1}^{n} prec(k_i).$$

The Mean Average Precision (MAP) is a widely accepted evaluation measure in Information Retrieval (Manning et al., 2008). The MAP measure is strongly correlated with query difficulty, which is reflected by the length of the ranked list and the size of the correct item set.

5.1.2 Normalized Discounted Cumulative Gain (nDCG)

Usually the list of recommended items is ranked from most to least relevant. When that is the case, a useful metric is the Discounted Cumulative Gain (Agichtein et al., 2006; Jarvelin & Kekalainen, 2000), which measures how effective the recommendation method is at locating the most relevant items at the top and the least relevant items at the bottom of the recommended list. This metric is based on human judgments. Human judges rate on how relevant each retrieval result is on an *n*-point scale. For a given recommendation list, the ranked results are evaluated from the top ranked down, and the *DCG* at a rank position *p* is computed as shown below:

$$DCG_p = rel_1 + \sum_{i=2}^{p} \frac{rel(i)}{\log_2(i)}$$

where rel_1, and *rel(i)* is a relevance rated by a user at the rank position 1, and one at position *i*. Each *rel(i)* is an integer representing the relevance rated by human judges (0 = "Not relevant to user interest at all" and 5 = "Perfect Relevant to user interest" at position *j*). Then, the nDCG is:

$$nDCG_p = \frac{DCG_p}{IDCG_p}$$

where IDCG is the ideal DCG of the perfect ordering list that is ranked in the most relevant items at the top, and lesser relevant items at the bottom. NDCG weighs relevant documents in the top ranked results more heavily than those ranked lower and punishes irrelevant documents by reducing their contributions to NDCG (Agichtein et al., 2006).

5.2 GENERAL CONSIDERATIONS FOR ALL EXPERIMENTS

5.2.1 Offline System

The data from 29 conferences, which were talks and users, were available in Conference Navigator 3 (CN3). Those conferences consist of Hypertext 2008, Adaptive Hypermedia 2008, Hypertext 2009, UMAP 2009, EC-TEL 2009, UMAP 2010, EC-TEL 2010, ASIS&T 2010, iConference 2011, UMAP 2011, Hypertext 2011, TPRC 2011, EC-TEL 2011, iConference 2012, UMAP 2012, Hypertext 2012, EC-TEL 2012, CSCW 2013, iConference 2013, LAK 2013, Hypertext 2013, UMAP 2013, AIED 2013, EC-TEL 2013, i-KNOW 2013, IUI 2014, UMAP 2014, Hypertext 2014, and EC-TEL 2014. The first three experiments were conduced using data from the first 23 conferences. The external validity was assessed using data from the last six conferences. They are described in the section 5.4.1.

5.2.2 Online System

The data of research talks, which are the talks themselves, and users were available in CoMeT. The experiments 4 and 5 that are described in section 5.4.2 and 5.4.3, respectively, were conducted using this system.

5.2.3 What is Being Recommended and to Whom

The offline system recommended conference talks. The long research and short research papers in the main conference, keynotes, posters, and workshop papers were considered as talks. The users of this recommender were ones who had provided bookmarks in the past.

The online system recommended research talks. The research talks were limited only talks that occurred in the Fall 2013 semester (September 1, 2013 – December 16, 2013) for study 4 and in the Spring 2013 semester (January 10 to February 5, 2013) for study 5. The subjects of this recommender were ones who had provided bookmarks or ratings in the past.

5.2.4 Clustering Approaches

The heuristic-based approach is used for clustering users to make the group-level user profile representations in studies 1 and 3. Three clustering algorithms were chosen. K-means, K-medoids, and hierarchical clustering were selected for the clustering approach. Bayesian Information Criterion (BIC) was used as an indicator to identify the quality of clusters. Every clustering approach was repeatedly conducted ten times in order to find the good quality of clusters. Hierarchical clustering approach was the first one to conduct clustering and used the elbow stop-criterion method to cut off the dendrogram in order to determine a number of clusters. K-Means and K-Medoids used the number of clusters from the hierarchical clustering result in their parameters.

5.2.5 Baselines

There are six baseline approaches (6 CBF + 2 CBCF), consisting of six from the content-based filtering recommender systems (CBF) and two from the collaborative filtering recommender systems (CBCF).

5.2.5.1 Content-Based Filtering Recommender System (CBF)

There were two dimensions for the content-based recommender systems in order to construct baselines: the user profiling approach, unigram vs. SVD; and the profile granularity, individual-level user model vs. cluster-level user model.

- **Individual Vector Space Profiles**: The single centroid vector space profiles contain the vector of unigram terms, which are extracted from the content of papers (title and abstract) users have bookmarked. The TF-IDF weighting is used as a scheme in order to weight terms with more frequency in the document (Term Frequency) but discount them if they appear too often in the corpus (Inverse Document Frequency). The user profiles are the average of every TF-IDF vector from documents user bookmarked (as previously mentioned in the section 2.3.1). The talk recommendation to users is the rank of cosine similarity between TF-IDF of user profiles and the target talks.
- **Individual Latent Semantic Vector Space Profiles**: The latent topics are extracted by using singular vector decomposition (SVD) from the TF-IDF unigram-document matrix as mentioned in the unigram vector space profiles. The SVD technique is applied to the matrix in order to get the *U* matrix. The profile for each user is an arithmetic mean of certain *k* latent topics from the *U* matrix regarding documents users bookmarked or rated.

The *k* number of latent topics starts from 100, stepping up 100 until the latent topics number is 2000.

- **Cluster Vector Space Profiles**: For cluster profile granularity, talks are considered to be included into the profiles if any member of the cluster bookmarked or rated.
- **Cluster Latent Semantic Profiles**: From the five-fold cross-validation result, the certain *k* latent topic is chosen as yielding the maximum result and it is carried on to use in the clustering profiles granularity. The same way as cluster vector space profile, talks are considered to be included into the profiles if any member of the cluster bookmarked and the number *k* of latent topics from the individual profile granularity to reduce the number of conditions in the study.
- **Positive-sample-only *k*-Nearest-Neighbors with Unigram Vector Space Model**: The approach is borrowed from the KNN regression. The recommending talks are ranked by their average cosine similarity to the *k* nearest talks from the user profile. The unigram terms TF-IDF vectors represent the talks.
- **Positive-sample-only *k*-Nearest-Neighbors with Latent Semantic Vector Space Model:** This approach works the same as the previous KNN.PO, one but this one uses the latent semantic vectors instead of unigrams.

5.2.5.2 Content-Boosted Collaborative Filtering Recommender System (CBCF)

In this recommender type, only one dimension, which is the user profiling approach: unigram vs. SVD was considered. In this thesis, user-based nearest neighbor collaborative filtering recommendations were used. The underlying assumption was that the users who had similar

preferences/tastes were more likely to select the similar items. The cosine similarity was chosen as the method to calculate similarity between the user and its peers:

$$Sim(u,v) = \frac{U \cdot V}{|U| \cdot |V|}$$

where *U*, and *V* are the TF-IDF vectors (unigram vector space or SVD ones) of user *u*, and user *v*, respectively. For the calculation of the recommendation, the *k* neighboring peers (*k*NP) were selected. In this dissertation, a number of nearest neighboring peers was set to ten for the CBCF recommendation. From the ten nearest peers, the recommended talks were selected from talks whose peers bookmarked or rated, but excluding ones that were bookmarked or rated by the user itself. The weighted score or predicted rating of these talks were calculated by a summation of production of cosine similarity between a user and its peers and their normalized rating on those talks (0 or 1 as in the case of bookmark), divided by a summation of a summation of cosine similarity between a user and its peers as shown in an equation below:

$$PredictedRating(u,i) = \frac{\sum_{v \in kNP}(Rating(v,i) \times Sim(u,v))}{\sum_{v \in kNP} Sim(u,v)}$$

where $Rating(v,i)$ is a function returning user rating (0 or 1 as in the case of bookmark) if user *v* bookmarked or rated document *i*, and $Sim(u,v)$ is the cosine similarity between user *u* and user *v*. Then, these talks were ranked based on their predicted ratings.

The way to calculate CBCF recommendation on the latent semantic (SVD) profiles is mostly the same as one of the unigram vector space profiles. The key difference is that the SVD profiles comprise the number of *k* latent topics instead of unigram TF-IDF vector. To keep a similar investigation manner as CBF on latent profiles, the *k* latent topics were chosen by starting

from 100, stepping up by 100, until 2000. The goal of this stepping exploration was to find the optimal *k* topics yielding the best result in the cross-validation.

5.2.6 Experimental Models with External Source Augmentation

There were three external sources taken into account in this dissertation: homepage, user publication (bibliography), and external scholarly bookmarked papers. As mentioned earlier in the baseline models, four CBF approaches and two CBCF approaches were conducted with augmenting external sources.

5.2.6.1 Content-Based Filtering Recommender System (CBF)

With external sources, the introduction of extra unigram terms is more likely to happen. These extra terms affect the vector space TF-IDF vectors. Including terms not existing in the original corpus would increase the sparseness of the vector space matrix and reduce the value of whereabouts of the terms from the original target content. As a result in order to explore this effect, the TF-IDF vectors built consist of two types of vectors:

1) TF-IDF vectors taking ONLY terms from the original corpus into account in both Term Frequency (TF) and Inverse Document Frequency (IDF) calculation, and
2) TF-IDF vectors considering terms from both the original target corpus and external ones.

These two types of TF-IDF vectors were applied to all external source augmentation models.

Similar to the baselines, four profiling representations: the individual vector space profile, the individual latent sematic (SVD) profile, the cluster vector space profile, and the cluster latent semantic (SVD) profile were in use.

- **Individual Vector Space Profiles** augmented with external sources are constructed on both TF-IDF vector types.
- **Individual Latent Semantic Profiles** (SVD) with external source augmentation were explored the same as the baseline profiles do, starting from 100 latent topics from both TF-IDF vector types, stepping up 100 until the latent topics number is 2000. Two types of TF-IDF vectors, excluding or including the terms not existing in the original corpus, were used in order to construct the user profiles.
- **Cluster Vector Space Profiles:** The target user profile were injected with external sources only, and the others profiles were constructed without any augmentation. Three clustering techniques were used in order to group these vector space user profiles. Then, the cluster vector space profile was constructed by averaging the talks that any member of the cluster bookmarked or rated along with the external source documents. Also, two types of TF-IDF vectors were used in order to construct the user profiles.
- **Cluster Latent Semantic Profiles** with external source augmentation used the *k* latent topics that produced the best result from the individual latent semantic profiles with external source augmentation as mentioned earlier. The individual latent semantic (SVD) user profiles were built and clustered with three techniques.

5.2.6.2 Content-Boosted Collaborative Filtering Recommender System (CBCF)

The scenario of CBCF with external source augmentation experiments, in this dissertation, was that, in each evaluation of a single user, only target user profiles were associated with the external information, while the profiles of peers did not use any external information. As a result, there was only one user profile that was enhanced with the external information at a single time of evaluation per user. The way to calculate the recommendation was also similar to CBCF baseline models. It was the user-based nearest neighbor collaborative filtering.

In the latent semantic SVD user profiles, the user profile of target user was constructed in the same way as the SVD CBCF baseline. The user profile was a latent semantic vector transformed from the external-source-augmented TF-IDF-weighed unigram vector. The most critical step was to choose a number of latent topics in the user profile. A number of *k* latent topics from the augmented latent semantic vector started from 100, stepping up by 100, until 2000 in the same way as in the baseline models. As results, there were twenty models for the excluding-extra-term TF-IDF vectors and twenty ones for the including-extra-term TF-IDF vectors.

5.3 PRELIMINARY CONCEPTS

Main Independent Variables: The independent variables were the recommending approaches with or without external source augmentation (Study 1, 2, 4.1, 4.2, and 5.1) and the combination of their recommendations (Study 3, 4.2, and 5.2). The Term Frequency/Inverse Document Frequency (TF-IDF) was used to compute the term weight for unigram profile

approach. For LSI-based approaches, the latent semantic matrix of TF-IDF vector was used as user profile representation. In this experiment, cosine similarity was used as a similarity ranking.

Main Dependent Variables: The dependent variables were MAP, the NDCG of Relevance, Attending, and Novelty of recommendation lists. Recommending research talks by the external-source-augmented approach for each recommender type (CBF or CBCF) was expected to provide better (higher) MAP, Relevance, Attending and Novelty NDCG than baseline (CBF or CBCF).

5.4 THE OVERVIEW OF EXPERIMENTAL DESIGN

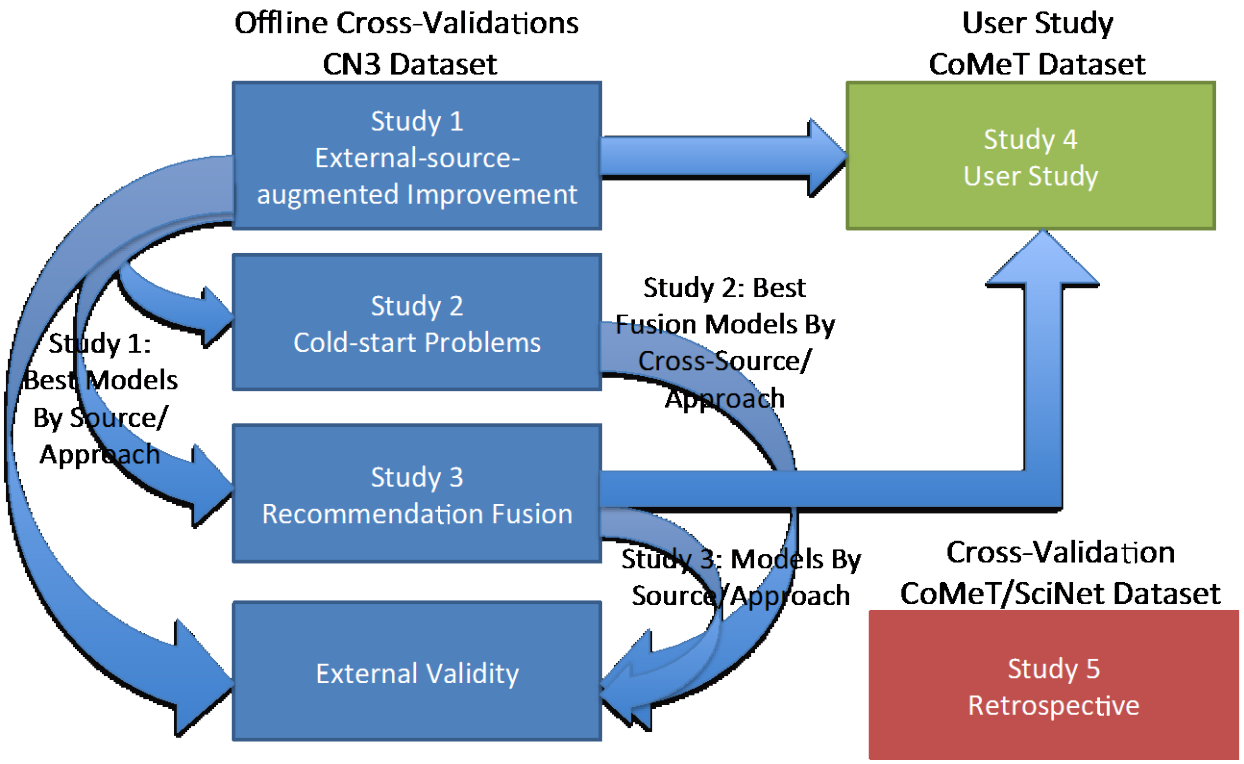


Figure 11: Study Flow

The flow of experiments from the offline study and the user study is shown in Figure **11**. Starting from the offline cross-validation, the best performing models based on external source and type of recommending method were carried to latter offline studies. The external validity was conducted to re-confirm the findings from the offline Study 1 – Study 3 with their most performing models. Moreover, best performing models from the offline Study 1 – Study 3 also were brought to Study 4 and Study 5.

5.4.1 Studies 1-3: Offline Cross-Validation on CN3 Dataset

The external source information, such as personal website, bibliography, and external bookmarks, for these offline validations was extracted from CN3 dataset and their content was crawled from their websites or services. In addition, users were matched by their names to find the external sources if they did not provide their information in the CN3 platform system.

5.4.1.1 Experimental Procedure

The talk information from Conference Navigator 3 platform regarding 23 conferences were obtained. There were 3321 research talks from those conferences in CN3 and 292 ones from the holdout for the external validity. There were 434 users in CN3, who had at least five papers and at least one external source that could be used in the experiments. The five-fold cross-validation assessment was conducted in all offline studies and external validity, which had talks that user bookmarked in the holdout dataset on the EC-TEL 2013, i-KNOW 2013, IUI 2014, UMAP 2014, Hypertext 2014, and EC-TEL 2014 conferences.

The qualified users in CN3 must have bookmarked at least five papers in any conference and provided at least one external source, consisting of:

1. The personal webpage (317 users),
2. Google Citation for the user publication source (170 users) and Scopus (79 users),
3. CiteULike (28 users) and Mendeley (17 users) accounts are provided in the CN3 user personal information.

If users published papers in the conferences CN3 had hosted before, the procedure mapping users with Scopus database was conduct. Matching was manually confirmed and those users were included in the pool of qualified users.

Five-fold Cross-validation: The five-fold cross-validation procedures were conducted by dividing bookmarked talks into five folds (bins) equally and randomly for every user. The evaluation ran five times for each user. For each run on each user, bookmarked talks from five folds were divided into two sets, training and test set. The training set contained a combination of bookmarked talks from four folds. Also, in the external-source-augmented models, the extra information from the external sources was included into the user profile. These talks from the training set and external user information were used for constructing the exploring recommendation approaches. The test set consisted of the talks from two sources. One was from the remaining fold, and another was from with the non-bookmarked talks left from the corpus. The evaluation result for each user was the average of five running results. Each study had a different detail, which is described in the chapter 6 through 9 respectively.

5.4.1.2 Browsing Log History Analysis Procedure

One of the contributions of this study on which user behaviors, such as user browsing logs, user bookmarks, and user ratings, in recommender systems, was the use of user browsing log analysis to understand in more detail the results of the hypotheses described above.

5.4.2 Study 4: User Study on CoMeT Dataset

The external source information, such as personal website, bibliography, and external bookmarks, for this CoMeT user study was provided by the subjects and their contents were crawled from their websites or services. Subjects were asked to bookmark research talks that they wanted to attend in the first part of the study. Afterward, subjects were asked to bookmarked talks and rate their relevance and novelty in the second part of the study.

5.4.3 Study 5: Retrospective Cross-Validation Study on CoMeT-SciNet Dataset

Data collection was performed at University of Helsinki, and the offline ten-fold cross-validation was conducted to evaluate the results. In the data collection process, subjects were asked to use SciNet as they were about to prepare for a course or seminar regarding to their research interest. SciNet has an exploratory search interface. Its open user model approach provides the ability to directly see a visual representation of the model and interact with it. The open user model is visualized as a radial layout. The radar provides the estimated search intent and its alternative intents represented by scientific keywords. Keywords are organized in the way that keywords relevant to the user are close to the center and similar intents have similar angles. More detail of

SciNet online user study procedure is elaborated in chapter 11. The external information from SciNet user models was extracted from the SciNet search logs. Later, subjects were asked to bookmark and rate research talks that they wanted to attend similar to study 4: user study on CoMeT Dataset.

6.0 STUDY 1: EXTERNAL-SOURCE-AUGMENTED RECOMMENDATION IMPROVEMENT

This chapter presents the effect of the external source to help improve research talk recommendations. The external source information, such as personal website, bibliography, and external bookmarks, for this experiment is extracted from CN3 dataset and their content is crawled from their websites or services. The results show that one of the selected external sources, bibliography, significantly helps improve the performance of content-based recommendation approach.

6.1 DATASET DEMOGRAPHICS

There are 3321 research talks from 23 conferences in CN3, consisting of Hypertext 2008, Adaptive Hypermedia 2008, Hypertext 2009, UMAP 2009, EC-TEL 2009, UMAP 2010, EC-TEL 2010, ASIS&T 2010, iConference 2011, UMAP 2011, Hypertext 2011, TPRC 2011, EC-TEL 2011, iConference 2012, UMAP 2012, Hypertext 2012, EC-TEL 2012, CSCW 2013, iConference 2013, LAK 2013, Hypertext 2013, UMAP 2013, and AIED 2013. 815 users in the dataset have at least five bookmarked talks. Among them, for 434 users in CN3 we identified at least one external source.

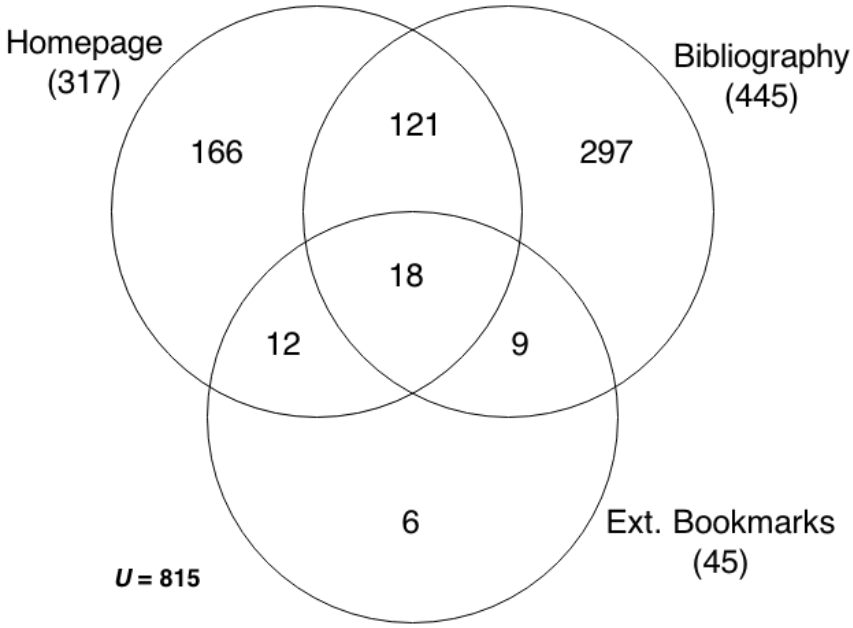


Figure 12: Users Demography in CN3 in Study 1

These 434 users were considered as qualified users in our study. The qualified users bookmarked at least five talks in any conference and also provided (or were matched to) at least one external source, consisting of: the personal webpage (317 users), user publication sources: Google Citation (170 users) and Scopus (79 users), and external scholarly bookmark sources: CiteULike (28 users) and Mendeley (seventeen users) accounts are provided in the CN3 user personal information.

The number of bookmarks per user follows the long tail distribution, ranging from five bookmarks to 729 ones. However, most of them have a small number of CN3 bookmarks. From the 434 qualified users, 343 of them (79.03%) have 5 to 30 bookmarks. Among these 343 users, 241 provide personal homepage (76.03% of users who provide homepage), 188 provide their publications (75.50% of users that we retrieved their bibliographies), and 26 provide the external bookmarked paper sources (57.78% of users who provide their CiteULike or Mendeley

accounts). In Figure **15**, there is a bump at 91 – 100 user publications due to the speed of the crawling process and limitations on Google servers not allowing massive downloading their data. As a result, the crawler was assigned not to download beyond 100 publications from Google Scholar.

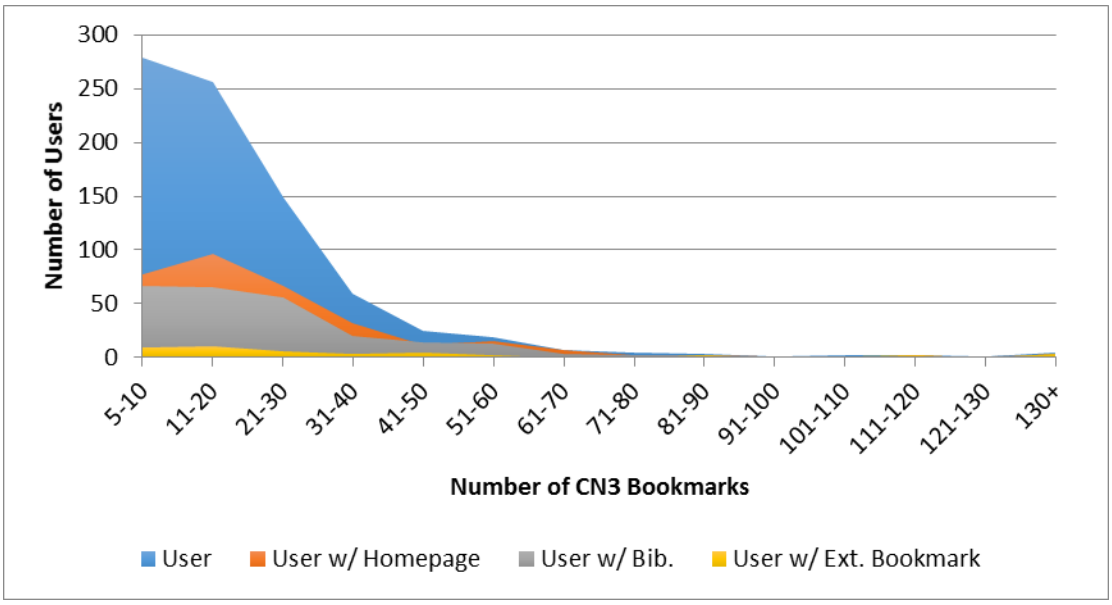


Figure 13: CN3 User Bookmark Distribution

260 out of 317 homepages that users provide (82.02%), have 10 pages or less. 149 out of 249 users (59.84%) whom we are able to identify and retrieve their publications, have 40 publications or less. However, 45 users have more than 90 publications (18.07%). Among 45 users who provide the external scholarly bookmark accounts, 27 out of them have 100 or less bookmarked papers in their accounts. 8 users have more than 450 bookmarked articles (17.78%) though.

The CN3 corpus has 10,410 unique unigrams. With external sources augmentation, the number of unique terms increases to 22,527 terms for homepage (29.29% overlapped), 16,394 terms for the external scholarly bookmark (50.52% overlapped), and 13,493 terms for bibliography (60.92% overlapped).

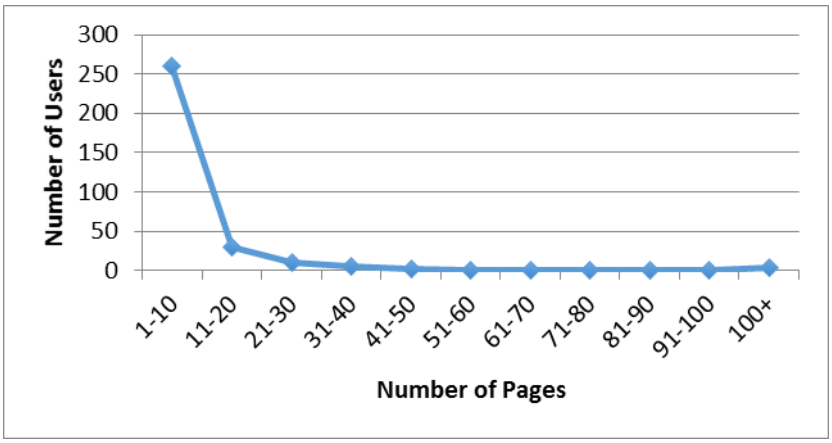


Figure 14: CN3 User Homepage Distribution

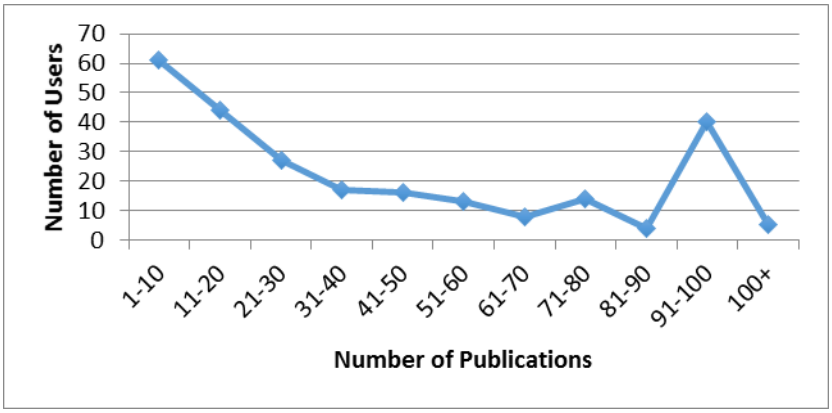


Figure 15: CN3 Bibliography Distribution

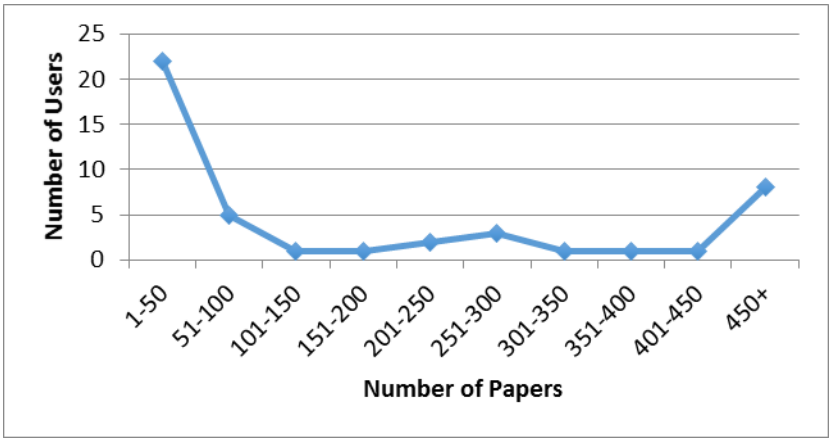


Figure 16: CN3 External Scholarly Bookmarked Papers Distribution

6.2 STUDY PROCEDURE

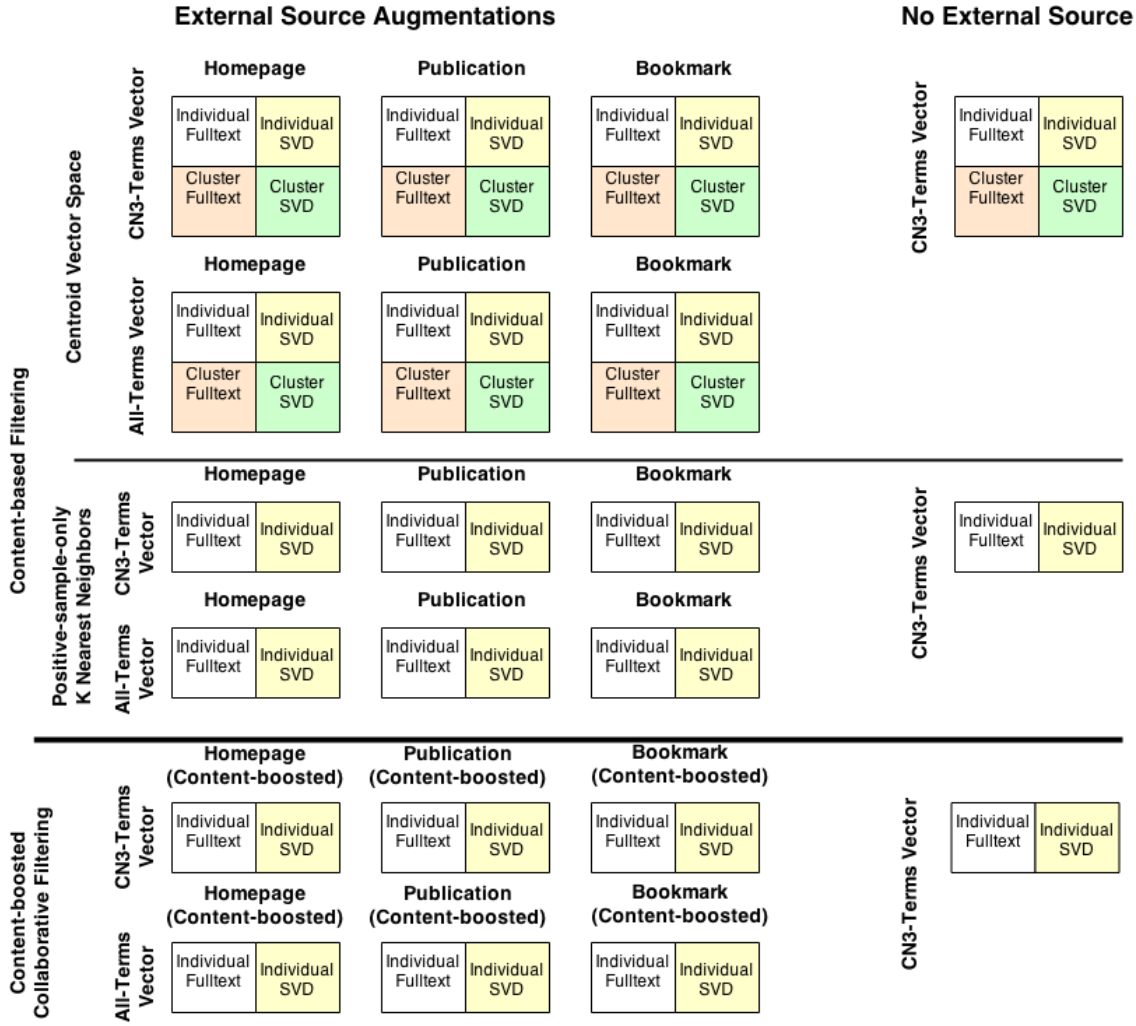


Figure 17: Step One: Generate the Candidacy Models Given Source and Recommender

In this study, the performance of different user profiling approaches are compared in the same external sources (homepage or user publication or external scholarly paper bookmarks, or none), for user profile representation (classical unigram vector space or SVD vector space), for user profile granularity (individual profiles or cluster profiles), for user profile application (Centroid

or KNN.PO), for each type of TF-IDF vectors (CN3-term TF-IDF vectors or All-term TF-IDF vectors) and for each recommender approach (Content-based Filtering (CBF) or Content-boosted Collaborative Filtering (CBCF)).

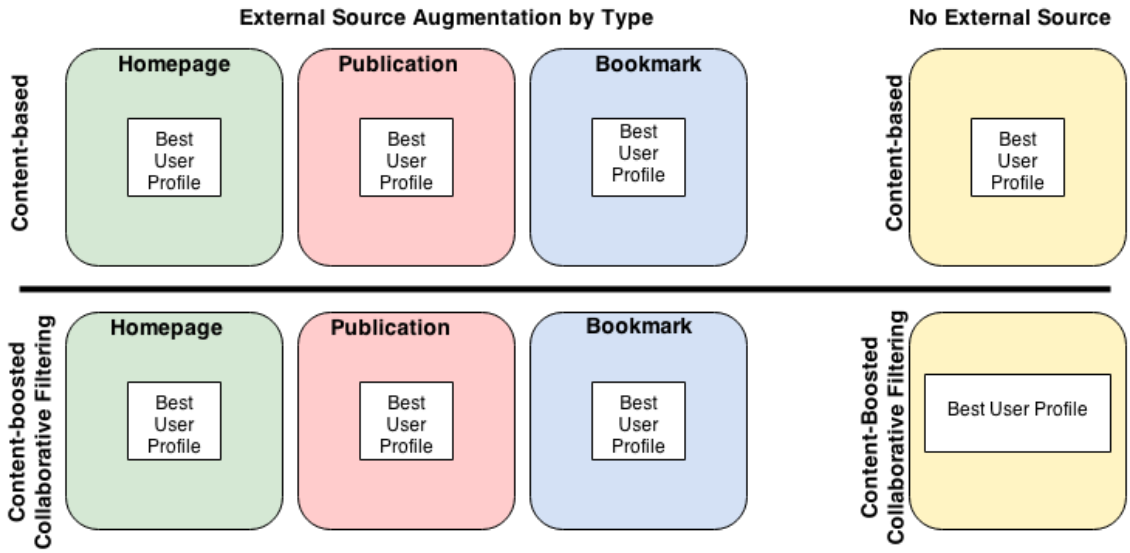


Figure 18: Step Two: Compare Those Candidates against the Baseline and Each Other.

Objective. The main purpose of this experiment is to discover the best recommendation approaches for each combination of external source and recommender approach. The best performing approaches selected at that stage are used in the next two studies (Study 2 and Study 3).

Setting. There are two steps on this study:

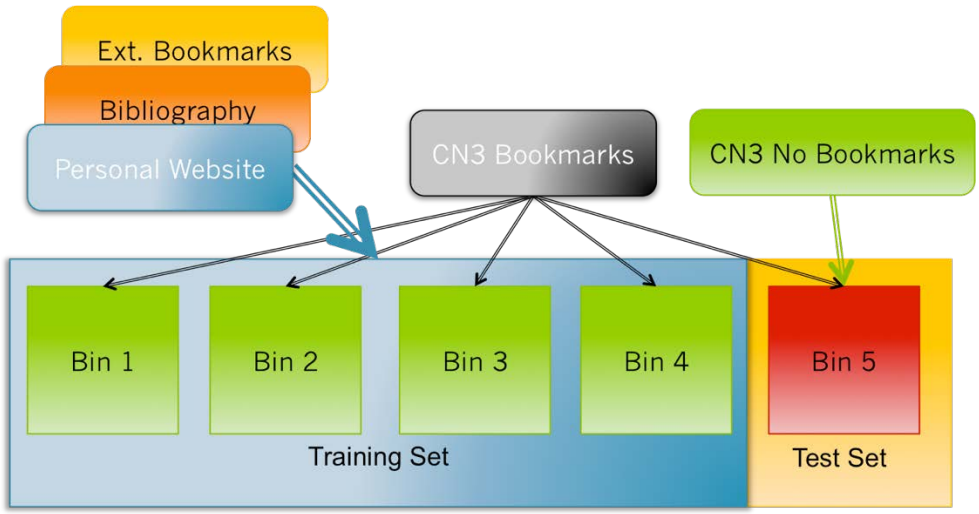


Figure 19: Five-Fold Cross-Validation

a) Step one is to generate candidacy models as depicted in Figure **17**. The reason is to reduce the complexity of the cross-validation study. The five-fold cross-validation procedure is mentioned earlier in section 5.4.1.1 and as depicted in Figure **19**. For evaluation of each user, the bookmarked CN3 talks are split into five folds (bins) equally and randomly. The evaluation runs five times for each user. For each run, bookmarked talks from five folds are separated into two sets, training and test set. The training set contains a combination of bookmarked talks from four folds. These talks are used for constructing the recommending approaches. Also, the external-source-augmented models include the extra information from the external sources into the training set. The test set consists of the talks from two sources. One is from the remaining one fold, and another is from with the non-bookmarked talks left from the corpus. The evaluation result for each user is the average of five testing assessments (one for each fold). In this step, the best user-profiling approach for each combination of source and recommender approach are selected.

b) Step two: candidate models are compared against the baselines.

Also, the study splits into two parts:

1. The comparison within the CBF approaches with or without external sources augmentations given any type of recommender systems,
2. The comparison within the CBCF approaches with or without external sources augmentations given any type of recommender systems.

Independent Variables: The independent variables are the recommending approaches with or without external source augmentation. The Term Frequency/Inverse Document Frequency (TF-IDF) was used to compute the term weight for unigram profile approach. For LSI-based approaches, the latent semantic matrix of TF-IDF vector was used as user profile representation. In this experiment, cosine similarity is used as a similarity ranking.

Dependent Variables. The dependent variable is the MAP. The recommendations returned from six experimental models with external sources augmentation (three CBF approaches + three CBCF approaches) are expected to provide higher MAP than the recommendation results returned from two best no-external-source baselines (CBF and CBCF).

Hypotheses of Study 1

Research Question 1

"*Which recommendation approaches and which external sources can deliver the best improvement over the traditional within-system recommendations?*"

Metrics: ***MAP***

H_0: There is no statistical difference between the *accuracy* means of recommended talks from CBF or CBCF,

	CBF	CBCF
Homepage	*with or without Personal Webpage augmentation.*	
User Publication	*with or without User Publication augmentation*	
Bookmarked Scholarly Papers	*with or without Bookmarked Scholarly papers augmentation*	

6.3 RESULTS

To assess the impact of external source augmentation, Study 1 was performed in two steps: (1) finding the best candidacy model on each source augmentation and each recommendation method, and (2) compare them with the baseline models. The results are analyzed in details below.

6.3.1 Baseline Recommendation with No External Source Augmentation

To understand some characteristics of recommendations without external source augmentation and before going into the experimental models with external source augmentation, the no-external-source-augmented models of CBF and CBCF, were assessed with all qualified 815 users and described in detail on the following sections.

6.3.1.1 Content-based Filtering (CBF)

There are three types of CBF models: individual centroid, clustering centroid, and Positive-Sample-Only *k* nearest neighbors (KNN.PO) models and for each model we have full-text and SVD user profile approaches. As mentioned in the previous chapter, the individual SVD CBF models were explored to find the optimum *k* latent topics that generated the maximum mean average precision (MAP) result as depicted in Figure **20**. The exploration showed that the MAP values started dropping after the number of latent topics reached 200. The maximum MAP result was 0.0573 for 200 latent topics. Therefore, the number of 200 latent topics was carried out to the other SVD CBF models, reducing the complexity of the experiment.

For the unigram clustering centroid models, as shown in Figure **21**, the hierarchical clustering, K-Means, and K-Medoids model did not produce good results. Their MAP results were 0.0303, 0.0304, and 0.0283 respectively. On the other hand, those from SVD clustering models returned MAP 0.0389, 0.0396, and 0.0348, that are significantly better than the same clustering methods with unigram vector space models (p-values: 0.00027, 0.0001, and 0.00163 respectively).

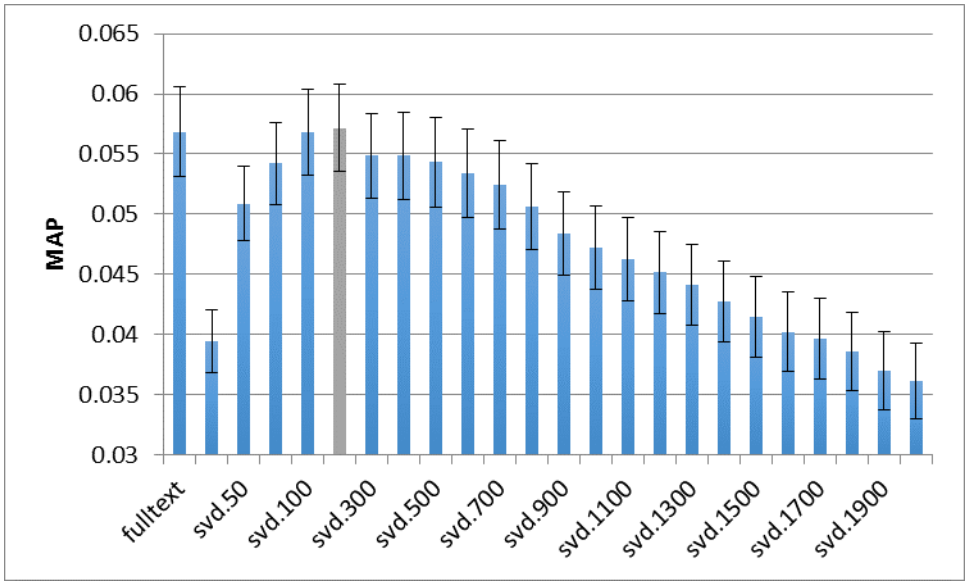


Figure 20: CBF Centroid Baseline MAP Comprised of Unigram Model and SVD Models

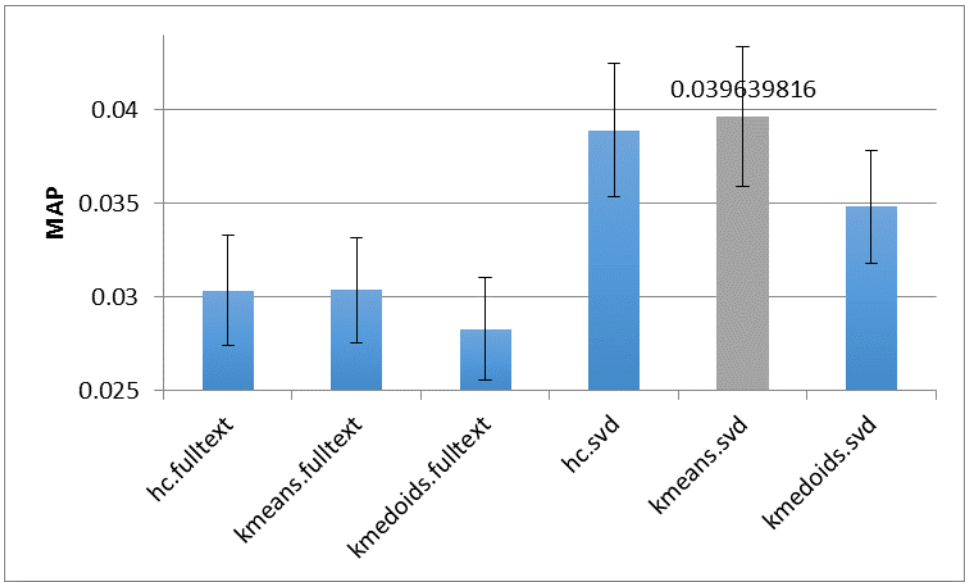


Figure 21: MAP results of Clustering Centroid Baselines

While two previous types of CBF models focus on the single user interest, the KNN.PO ones allow the models to capture multiple users' interests based on the closest average similarity

of k neighbors. The target talks were classified by ranking them with their average cosine similarity from the *k* nearest bookmarked talks from the user profile. The MAP results of KNN.PO models using the unigram terms and latent topics vector space models were shown in Figure **22**. For both KNN.PO models the MAP values started a bit low then gradually increased and peaked at the 15-NN.PO SVD model. The best MAP result was 0.0567.

The comparison among three top models from each CBF model in Figure **23** showed that the individual centroid models - latent semantic vector space model and KNN.PO SVD model - performed significantly better than the group K-Means SVD model (p-values: 8.30E-08 and 1.53E-07 respectively). However, the individual SVD model performed slightly higher than the individual KNN.PO SVD model.

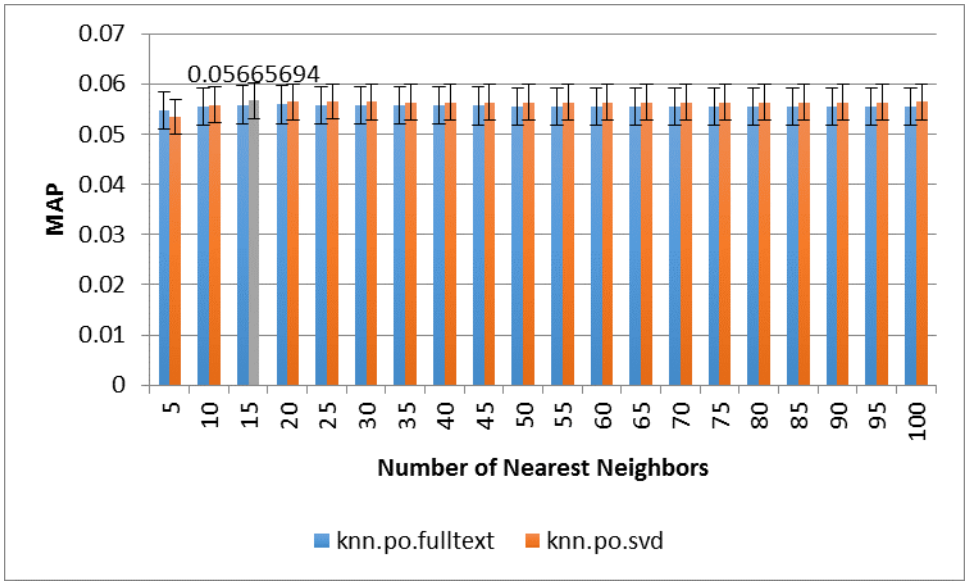


Figure 22: MAP Results of KNN.PO Baselines

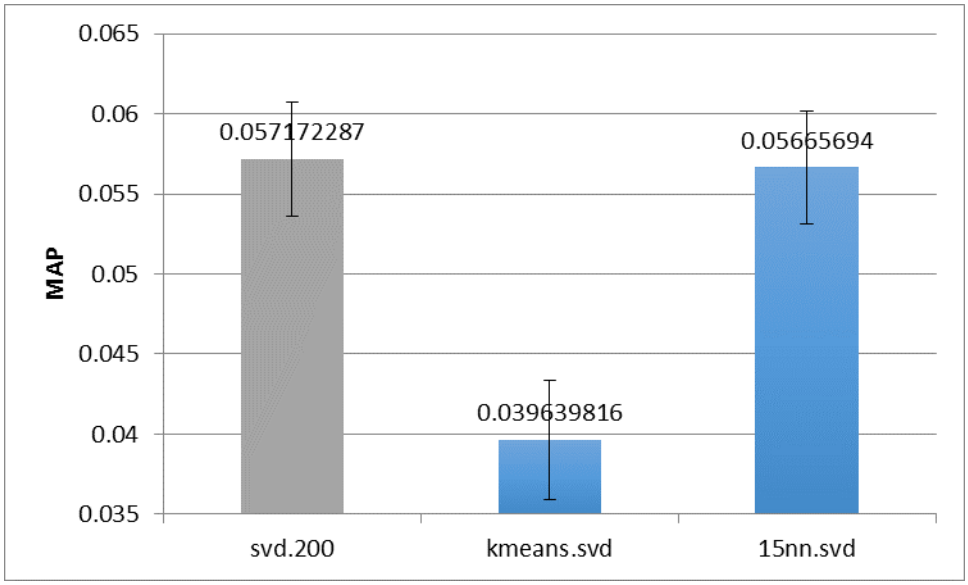


Figure 23: MAP Results of CBF Baselines

6.3.1.2 Content-boosted Collaborative Filtering (CBCF)

With the two types of TF-IDF vectors we used, there are two sub types of user-based nearest neighboring content-boosted collaborative filtering models: unigram vector space CBCF, and latent semantic (SVD) CBCF. The SVD CBCF models were explored, in a similar way as the CBF SVD baseline models, to find the optimum *k* latent topics that generated the maximum mean average precision (MAP) result as depicted in Figure **21**.

The MAP results from the SVD CBCF baseline models started the low MAP, getting better, later staying flat. The MAP peaked at the model with 1600 latent topics. The comparison between the unigram CBCF and 1600-latent-topic CBCF models showed that the SVD CBCF model performed slightly better than the unigram one but not significantly.

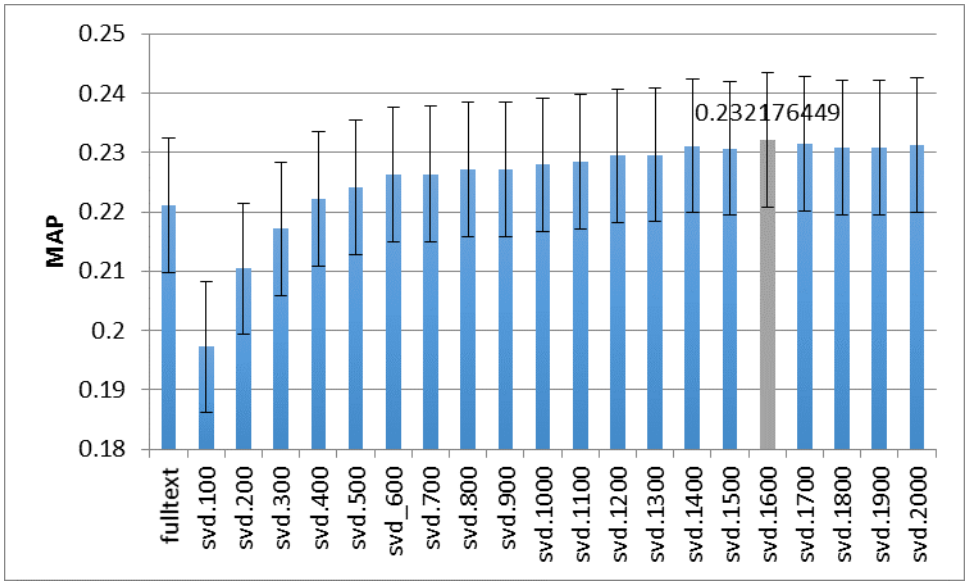


Figure 24: MAP Results of CBCF Baselines

6.3.2 External Source Augmentation: Homepage

The experimental models with homepage augmentation for CBF and CBCF were assessed with 317 users for which their homepages were either provided or identified in CN3.

6.3.2.1 Content-Based Filtering

(a) Homepage-Augmented CBF

There are three main types of homepage-augmented CBF models: individual centroid vector space, clustering centroid, and KNN.PO models. Even though, unlike the baseline, homepage and two other external source augmentation models have two TF-IDF vectors, one having only CN3 terms, and another one having terms from both CN3 and the specific external source.

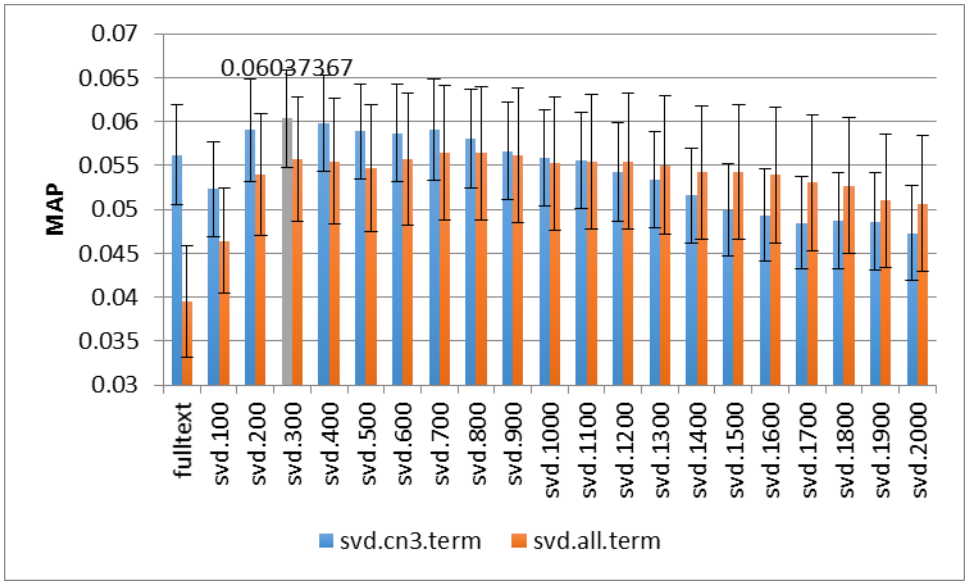


Figure 25: MAP Results of Individual Homepage Centroid Models

The individual homepage-augmented centroid CBF models on both TF-IDF vectors were explored to find the model that generated the maximum mean average precision (MAP) result as depicted in Figure **25**. The result of CN3-term SVD exploration was the MAP values starting with low MAP, going up, staying flat and dropping after the number of latent topics was more than 300. The maximum MAP result was 0.0604 at the 200 latent topics. The All-term SVD exploration looked a bit more like a plateau, starting from a low MAP, then staying steady, and starting going down after the 700 latent topics. As shown Figure **25**, the maximum MAP result was 0.06037 at the CN3-term SVD model with 300 latent topics. As a result, this model was selected as a representative of homepage-augmented individual centroid models. Also, the number of 300 latent topics was later used in the other homepage-augmented SVD CBF models.

For the full-text clustering centroid models, as shown in Figure **26** for CN3-term TF-IDF vectors and All-term TF-IDF vectors, the hierarchical clustering, K-Means, and K-Medoids

model did not produce good results. Their MAP results were 0.0303, 0.0304, and 0.0283 respectively. The cluster latent semantic models all used the 200 latent topics that were carried out from the individual SVD model. On the other hand, those from cluster latent semantic (SVD) models, 0.0313, 0.0351, and 0.0339 for CN3-term vectors and 0.0301, 0.0310, and 0.0253 for All-term vectors, were significantly better than the results from the same clustering methods with unigram vector space models (p-values: 3.89E-15, 2.22E-16, and 0 for CN3-term vectors and 0, 0, and 0 for All-term vectors respectively). As a result, K-Means CN3-term SVD model was selected as representative of homepage-augmented clustering centroid models.

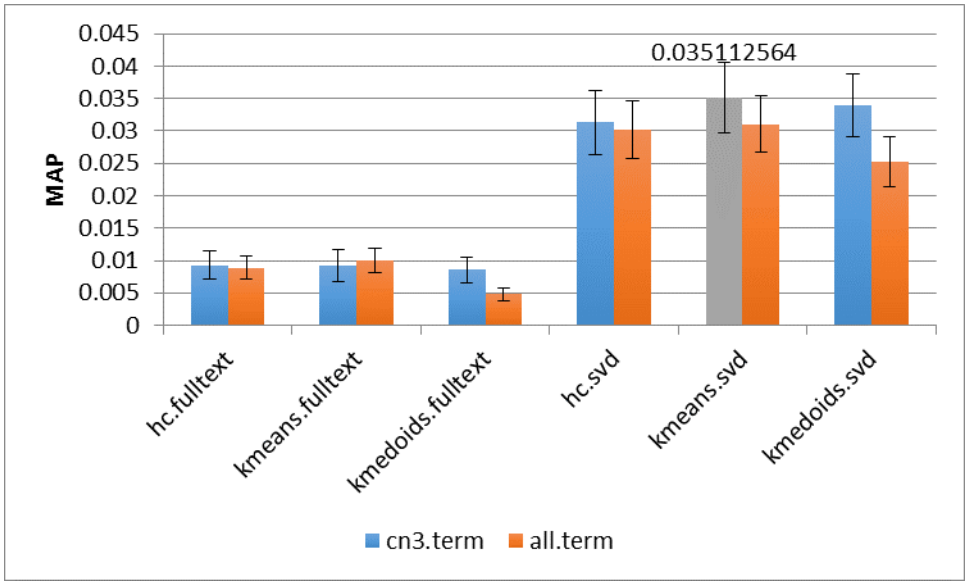


Figure 26: MAP Results of Homepage Cluster Models

The MAP results of homepage-augmented KNN.PO models using the full-text and latent topics vector space models with CN3-term and All-term vectors were shown in Figure **27**. The results of KNN.PO models with CN3-term vectors performed better than KNN.PO models with

All-term vectors on both unigram and SVD ones. Moreover, the CN3-term full-text KNN.PO ones significantly outperformed all of the All-term full-text KNN.PO ones. The maximum MAP result of all KNN.PO models was at the 20-NN.PO CN3-term SVD model. The maximum result was 0.062. As a result, the 20-NN.PO CN3-term SVD model was chosen as representative of the homepage-augmented KNN.PO models.

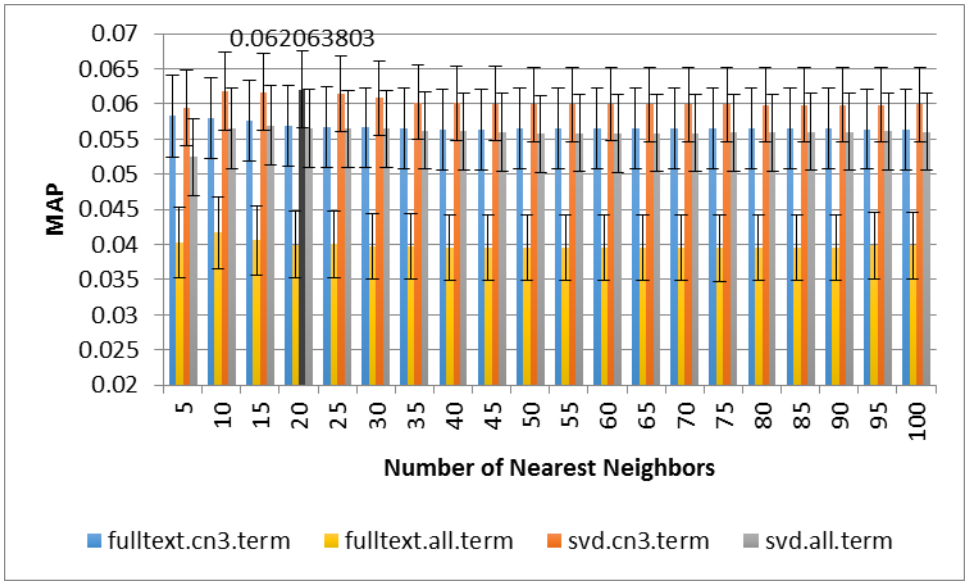


Figure 27: MAP Results of Homepage KNN.PO Models

The final comparisons in Figure **28** showed that the individual CN3-term SVD and 20-NN.PO CN3-term SVD models outperformed the K-Means CN3-term SVD model significantly (p-value: 1.73E-07 and 1.87E-08 respectively). However, the 20-NN.PO CN3-term SVD model performed slightly better than the CN3-term SVD model but not significantly. In this step, the 20-NN.PO CN3-term SVD model with 300 latent topics was chosen as the final Homepage CBF model.

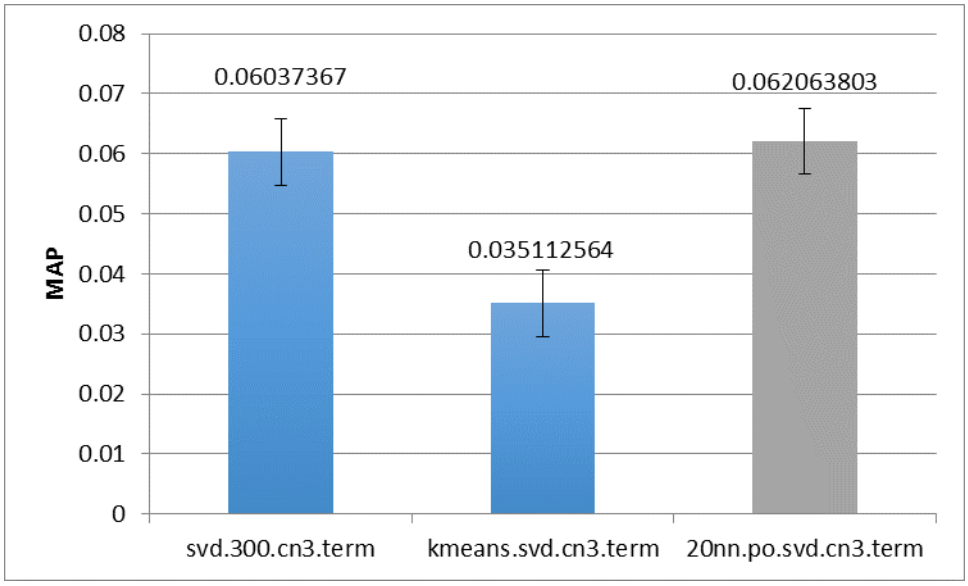


Figure 28: MAP Results of Homepage CBF Models

(b) Baseline CBF

In order to compare CBF baseline models with homepage-augmented CBF one from the previous section, the CBF baseline models needs to be re-evaluated with the same users. Therefore, the CBF baselines with the 317 users, for which their homepages were either provided or identified in CN3, were assessed as shown in **Figure 29**. The maximum MAP result of individual centroid models was 0.06076 on the SVD model with 100 latent topics which was the peak MAP result compared to the rest of them and the unigram vector space model as well, as shown in the top left diagram on **Figure 29**. Also, the number of 100 latent topics was later used in the clustering SVD centroid and KNN.PO SVD models.

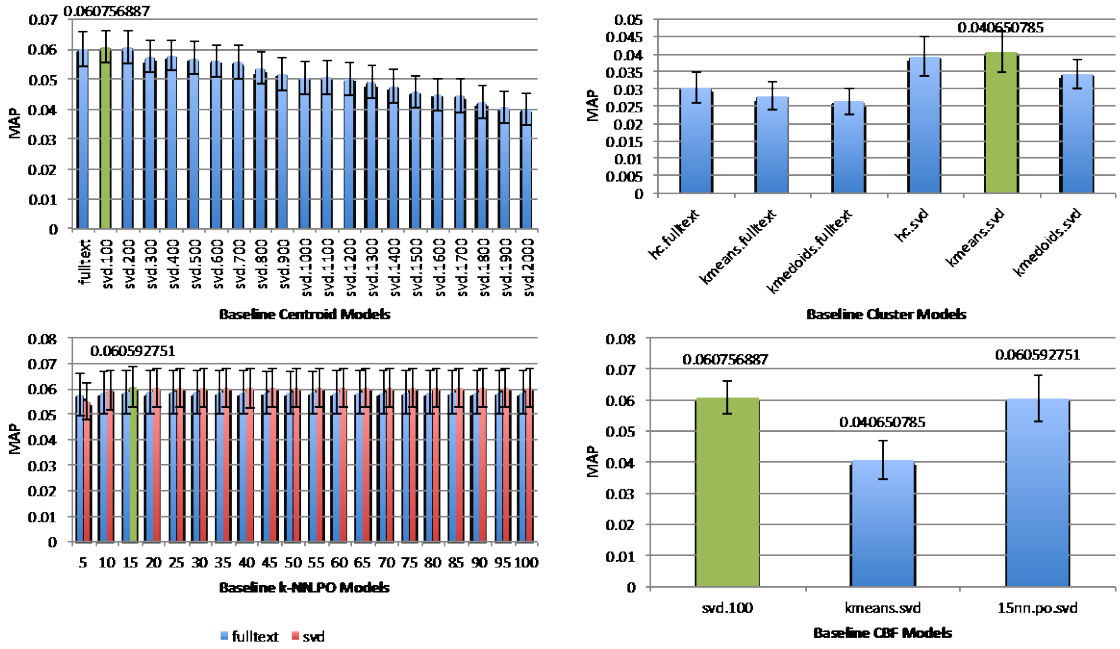


Figure 29: MAP Results of CBF Baselines with Homepage Users

All three clustering centroid SVD models performed better than the unigram clustering centroid models as seen on the top right diagram on **Figure 29**. The MAP result of K-Means SVD model was highest among the clustering models.

All the KNN.PO models performed similarly with no significant difference among them in the bottom left of **Figure 29**. The maximum MAP of KNN.PO model was 0.06059 on the 15-NN.PO SVD model. The final comparison showed that the individual centroid SVD model with 100 latent topics and the 15-NN.PO SVD model outperformed the K-Means SVD model significantly, but the result of the centroid SVD model was a bit higher than the KNN.PO one. As a result, the individual SVD model with 100 topics was chosen to test the hypothesis.

(c) Hypothesis Testing

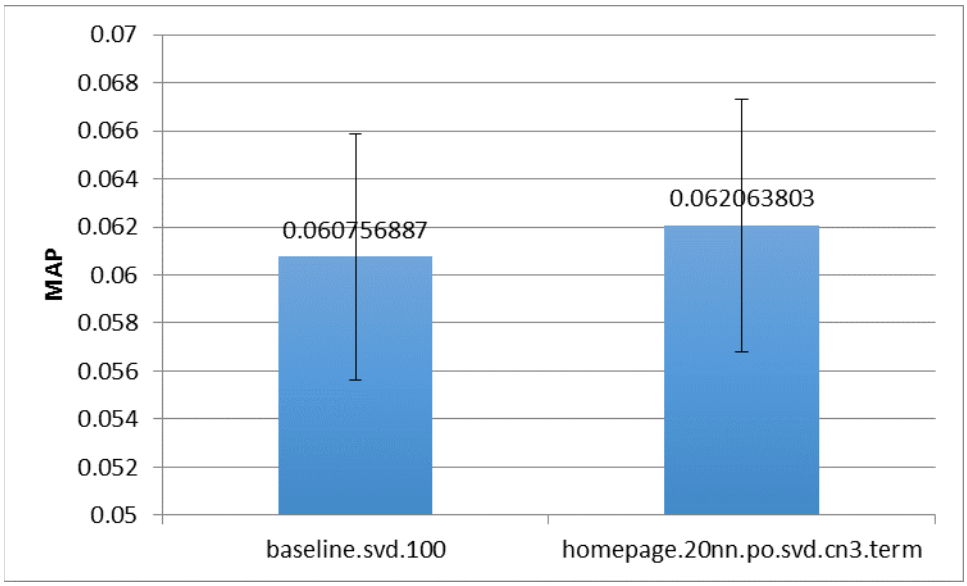


Figure 30: MAP Results of Homepage CBF vs. Baseline CBF

The homepage-augmented 20-NN.PO CN3-term SVD CBF model, with 300 latent topics, was assessed with 317 users, who provided their homepages or had these identified by us, against the SVD CBF baseline, with 100 latent topics.

H_0: There is no statistical difference between the *accuracy* means of recommended talks from any CBF, *with or without Personal Webpage augmentation*.

One-way Analysis of Variance (ANOVA) was applied to test the hypothesis. The following figure shows the results.

Analysis of Variance (One-Way)

Summary

Groups	*Sample size*	*Sum*	*Mean*	*Variance*
baseline.CBF.MAP	1585	96.29967	0.06076	0.01146
homepage.CBF.MAP	1585	98.37113	0.06206	0.01211

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*p-level*	*F crit*
Between Groups	0.00135	1	0.00135	0.11489	0.73467	5.41738
Within Groups	37.3262	3168	0.01178			
Total	37.32755	3169				

Figure 31: Hypothesis Testing Result on Homepage CBF Model

From Figure **31**, the null hypothesis is **accepted**. The mean of the results from the homepage CN3-term 20-NN.PO SVD model with 300 latent topics is statistically not different from the mean from the SVD CBF baseline with 200 latent topics at a p-value < 0.05 significant level.

6.3.2.2 Content-boosted Collaborative Filtering

(a) Homepage-Augmented CBCF

With two TF-IDF vectors for homepage source, there are four sub types of user-based nearest neighboring content-boosted collaborative filtering models: unigram CN3-term vector space CBCF, unigram All-term vector space, CN3-term latent semantic (SVD), and All-term latent semantic (SVD) CBCF.

As mentioned in the CBF SVD models, the SVD CBCF models were explored to find the optimum *k* latent topics that generated the max mean average precision (MAP) result as depicted

in Figure **32.** The MAP results from the CN3-term SVD CBCF homepage models started with a low MAP, getting better, and later staying flat. The MAP result for CN3-term SVD models peaked at the model with 700 latent topics. The All-term SVD models showed the same pattern that starting from a low MAP, going up rapidly, and then staying steady like a plateau. The MAP result for All-term SVD models was at a peak at the model with 900 latent topics.

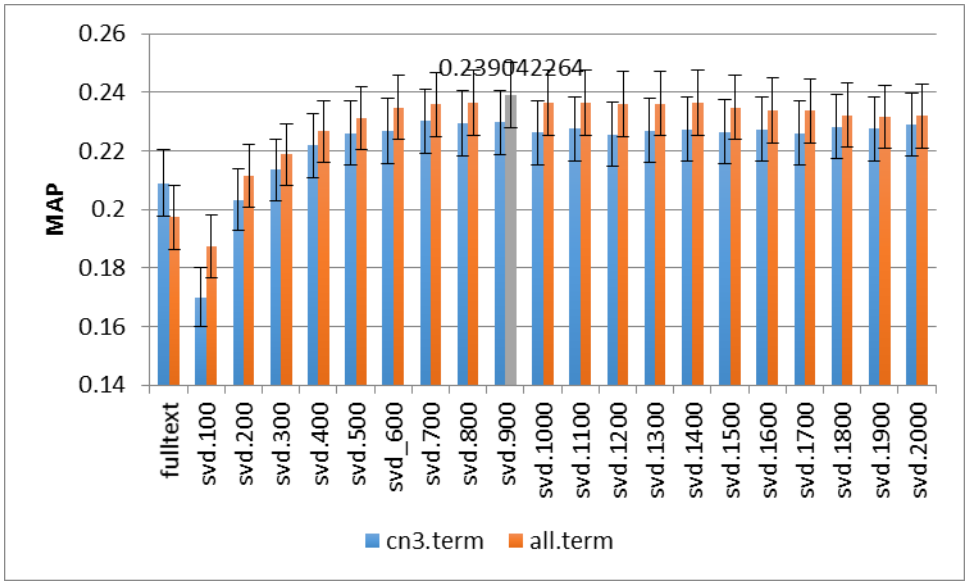


Figure 32: MAP Results of Homepage CBCF

The CBCF comparison of the unigram vector space models showed that the CN3-term CBCF is slightly better than the All-term unigram CBCF, however the latent semantic model, All-term SVD CBCF performed better than the CN3-term SVD CBCF one. Finally, the maximum MAP was obtained from the All-term SVD CBCF model with 900 latent topics and was selected as the homepage-augmented CBCF representative model.

(b) Baseline CBCF

With the same reasoning for CBF baseline, CBCF baseline needs to be re-evaluated with the same users with homepage-augmented CBCF models evaluated with only those users whose homepages are included. The CBCF baseline models with 317 users for whom we have been provided or identified homepages, were assessed as depicted in Figure **33**. Their MAP results stayed steady and the maximum MAP of CBCF baseline models came to the CBCF SVD one with 1600 latent topics.

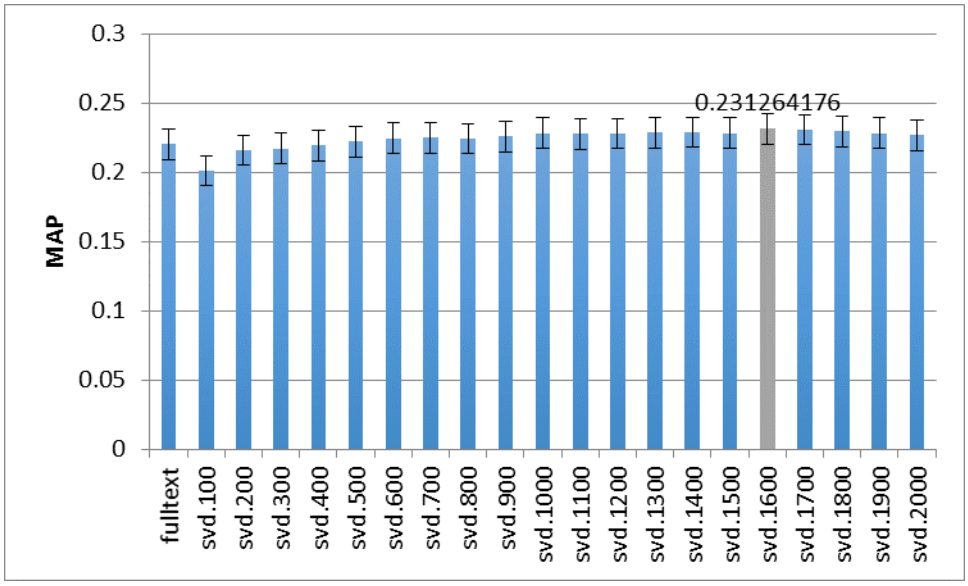


Figure 33: MAP Results of CBCF Baselines with Homepage Users

(c) Hypothesis Testing

The individual All-term SVD CBCF model with 900 latent topics with homepage augmentation was assessed with 317 users who provided their homepages with CN3 against the SVD CBCF baseline with 1600 latent topics.

H_0: There is no statistical difference between the *accuracy* means of recommended talks from CBCF, *with or without Personal Webpage augmentation*.

One-way Analysis of Variance (ANOVA) was applied to test the hypothesis. The following figure shows the results.

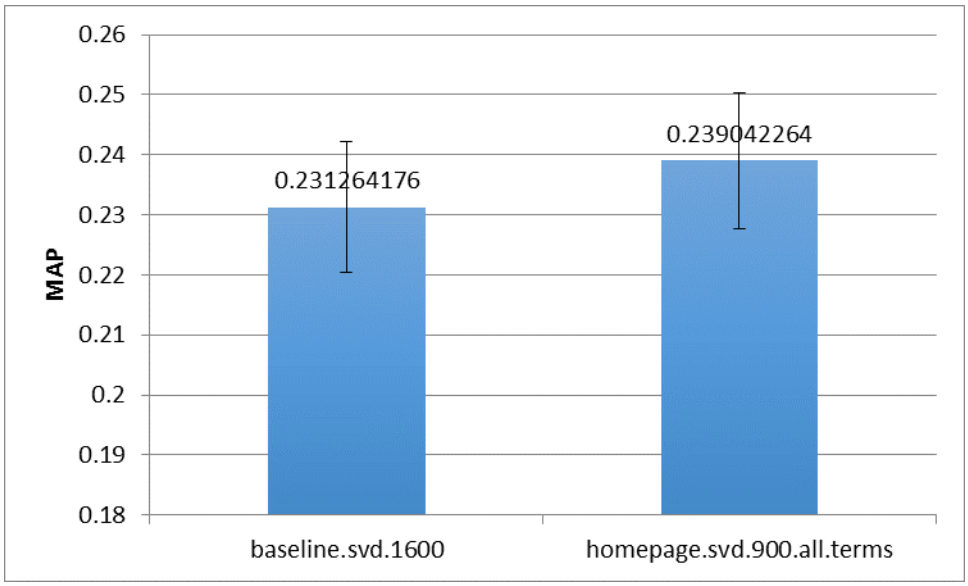


Figure 34: MAP Results of Homepage CBCF vs. Baseline CBCF

Analysis of Variance (One-Way)

Summary

Groups	*Sample size*	*Sum*	*Mean*	*Variance*
baseline.CBCF.MAP	1585	366.55372	0.23126	0.04896
homepage.CBCF.MAP	1585	378.88199	0.23904	0.05264

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*p-level*	*F crit*
Between Groups	0.04795	1	0.04795	0.94378	0.33138	5.41738
Within Groups	160.93912	3168	0.0508			
Total	160.98706	3169				

Figure 35: Hypothesis Testing on Homepage CBCF Model

From Figure **35**, the null hypothesis is **accepted**. Even though the mean of the results from the homepage All-term SVD CBF model with 900 latent topics was a bit higher than the mean from the SVD CBCF baseline with 1600 latent topics, it was not statistically different at a p-value < 0.05 significant level.

6.3.3 External Source: User Publication (Bibliography)

The experimental models with user publications (bibliography) augmentation for CBF and CBCF were assessed with 249 users for whom we retrieved their publications.

6.3.3.1 Content-based Filtering

(a) Bibliography-Augmented CBF

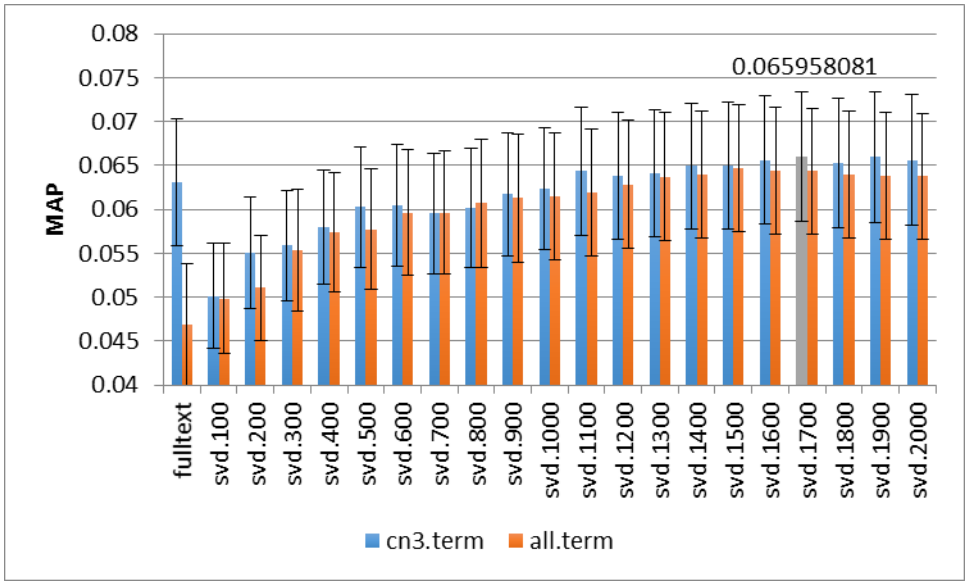


Figure 36: MAP Results of Bibliography Centroid Models

Like the homepage source, there are three main types of CBF models: individual centroid vector space, clustering centroid, and KNN.PO models. The individual centroid models on both TF-IDF vectors were explored to find the model that generated the maximum mean average precision (MAP) result as depicted in Figure **36**. The result of CN3-term SVD exploration was that the MAP values started with a low MAP, going up, and then staying flat. The All-term SVD models performed quite the same as the CN30-terms ones did. The maximum MAP result of individual all-term SVD models was 0.065 at 1500 latent topics. The maximum MAP result of the individual centroid models was 0.066 at the 1700 latent topics. As a result, the individual CN3-

term SVD model at 1700 topics was chosen as the representative of bibliography-augmented individual centroid CBF models.

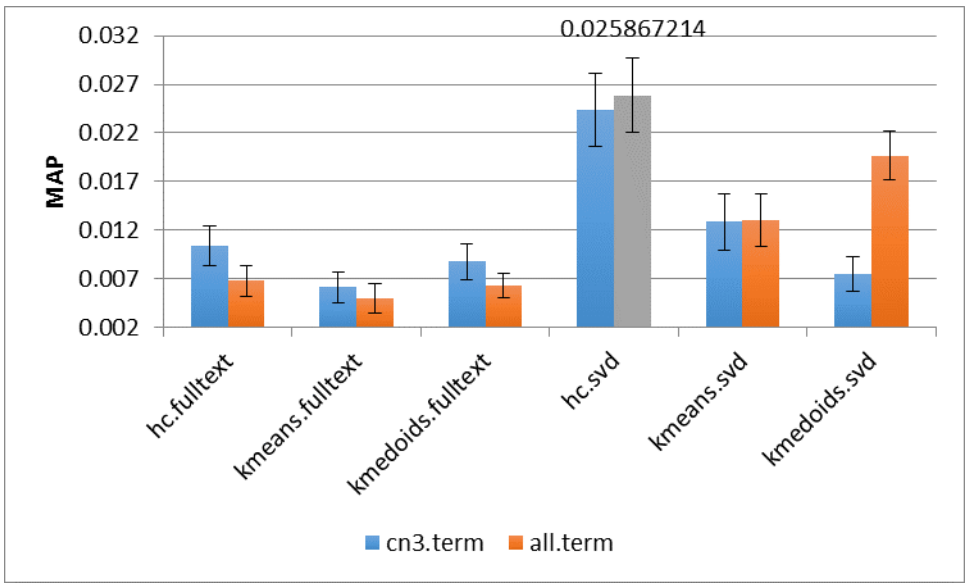


Figure 37: MAP Results of Bibliography Cluster Models

For both the full-text and the SVD clustering centroid models, as shown in Figure **37** for CN3-term TF-IDF vectors and All-term TF-IDF vectors, did not produce good results except the SVD hierarchical clustering models on both two TF-IDF vectors, the All-term SVD K-Medoids model. The cluster SVD models used the 1700 and 1500 latent topics that were carried from the individual SVD models for the CN3-term and All-term vectors respectively. From Figure **37**, hierarchical clustering SVD models with All-term vectors were selected as the representative of bibliography-augmented clustering centroid models.

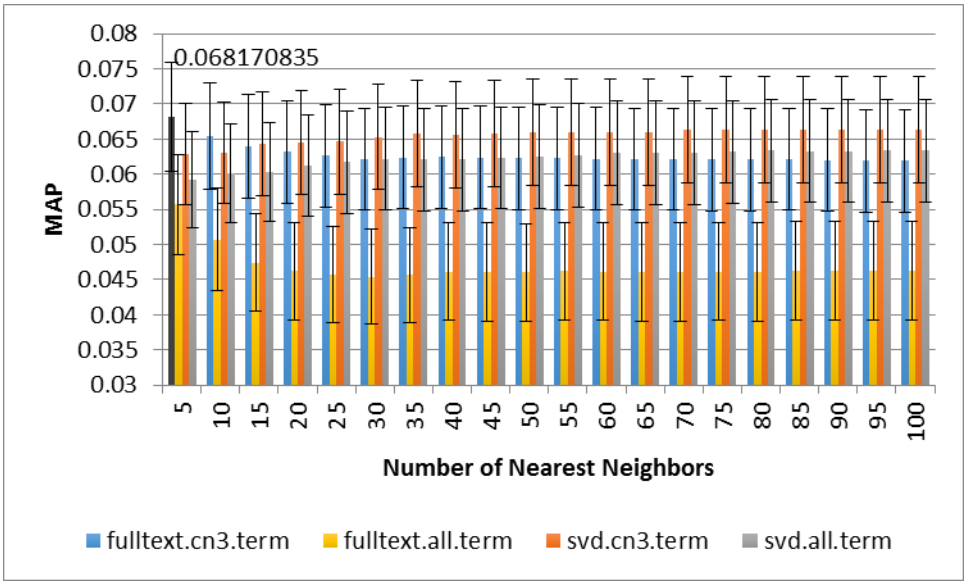


Figure 38: MAP Results of Bibliography KNN.PO

The MAP results of bibliography-augmented KNN.PO models using the unigram and latent topics vector space models with CN3-term and All-term vectors were shown in Figure **38**. The results of the bibliography-augmented KNN.PO models with CN3-term vectors performed slightly better than bibliography-augmented KNN.PO models with All-term vectors on both full-text and SVD ones. The maximum MAP result of all bibliography-augmented KNN.PO models was at the 5-NN.PO CN3-term full-text model. The maximum result was 0.0682. As a result, the 5-NN.PO CN3-term full-text model was chosen as the representative of the bibliography-augmented KNN.PO models.

The final comparisons in Figure **39** showed that the individual CN3-term SVD centroid and 5-NN.PO CN3-term unigram models outperformed to the All-term SVD hierarchical clustering model significantly (p-value: 4.22E-15 and 1.88E-15 respectively). The 5-NN.PO CN3-term full-text model performed slightly better than the CN3-term SVD centroid model but

not significantly. In this step, the bibliography-augmented 5-NN.PO CN3-term full-text model was chosen as the final bibliography-augmented CBF representative model.

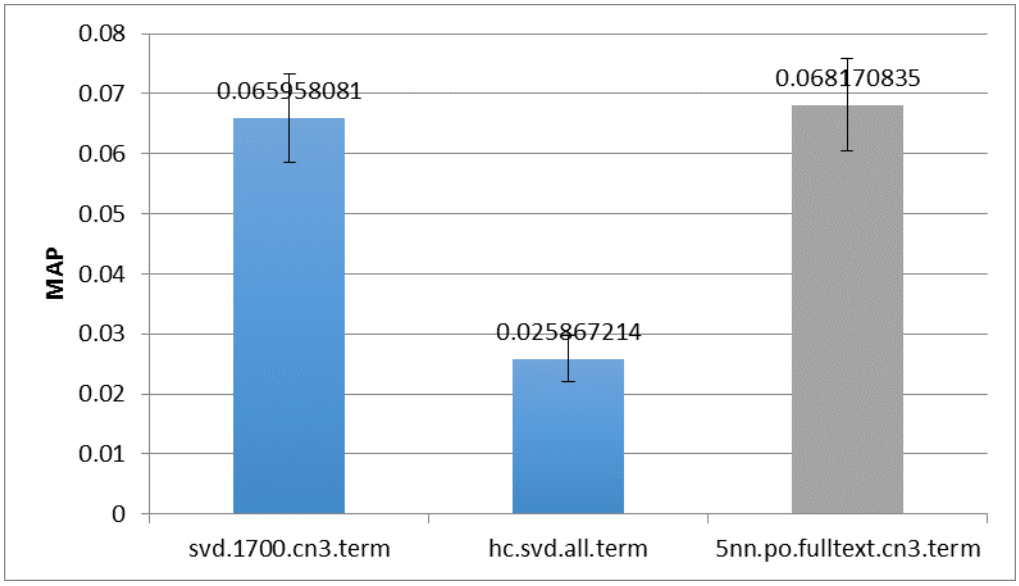


Figure 39: MAP Results of Bibliography CBF Models

(b) Baseline CBF

The CBF baselines with the same 249 users, who provided or for whom we identified their publications, were assessed as shown in Figure **40**. The maximum MAP result of individual centroid models was 0.0554 on the SVD model with 200 latent topics which was the MAP result peak compared to the rest of them and the unigram vector space model, as well, as shown in the top left diagram on Figure **40**. Therefore, the number of 200 latent topics was used later on the other bibliography-augmented SVD CBF models.

All the three SVD clustering models performed better than the unigram clustering centroid models as seen on the top right diagram on Figure **40**. The MAP result of K-Means SVD model was highest among the clustering models.

All the KNN.PO models performed similarly with no significant difference among them in the bottom left of Figure **40**. The maximum MAP result of KNN.PO model was 0.0554 on the 15-NN.PO SVD model.

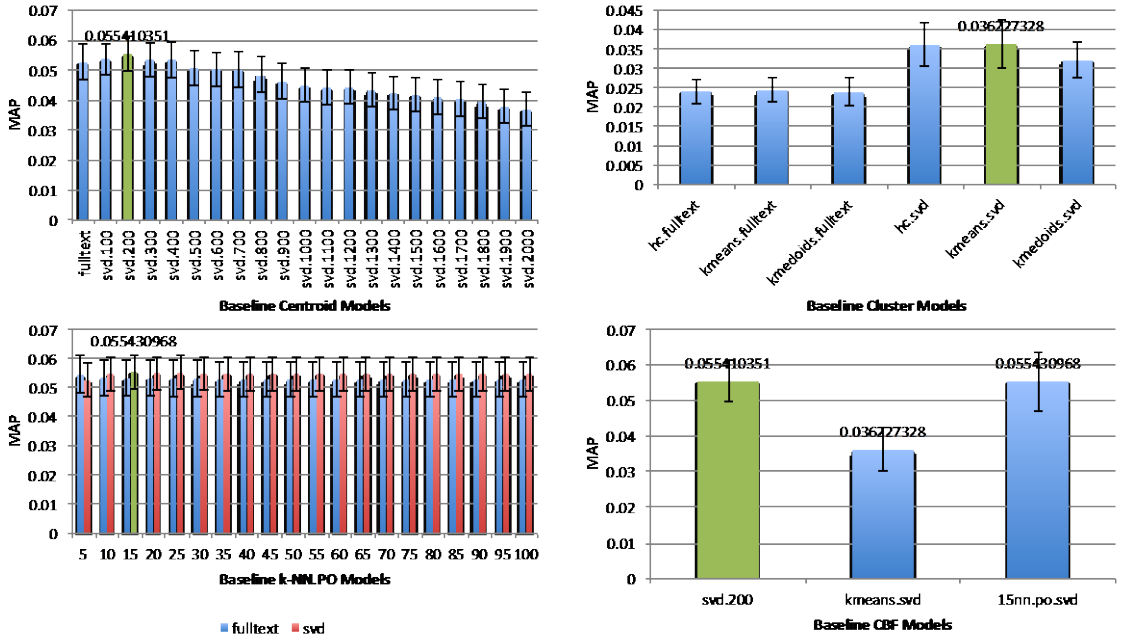


Figure 40: MAP Results of CBF Baseline Models with Bibliography Users

The final comparison showed that the individual centroid SVD model with 200 latent topics and the 15-NN.PO SVD model outperformed the K-Means SVD model significantly but the result of centroid SVD model was a bit higher than the KNN.PO one. As a result, the individual SVD model with 200 topics was chosen to test the hypothesis.

(c) Hypothesis Testing

The 5-NN.PO bibliography-augmented CN3-term full-text model was assessed with 249 users for whom we retrieved their publications against the SVD CBF baseline with 200 latent topics.

H_0: There is no statistical difference between the *accuracy* means of recommended talks from CBF, *with or without User Publication augmentation*.

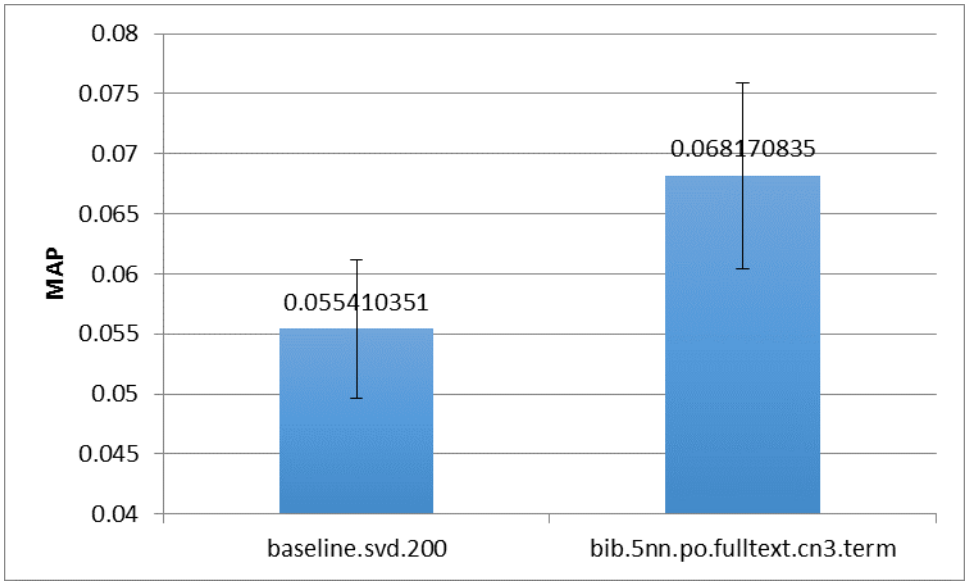


Figure 41: MAP Result of Bibliography CBF vs. Baseline CBF

One-way Analysis of Variance (ANOVA) was applied to test the hypothesis. The following figure shows the results.

Analysis of Variance (One-Way)

Summary

Groups	*Sample size*	*Sum*	*Mean*	*Variance*
baselineCBFMAP	1245	68.98589	0.05541	0.01089
bibCBFMAP	1245	84.87269	0.06817	0.01932

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*p-level*	*F crit*
Between Groups	0.10136	1	0.10136	6.70998	0.00964	5.41887
Within Groups	37.584	2488	0.01511			
Total	37.68536	2489				

Figure 42: Hypothesis Testing Result on Bibliography CBF Model

From Figure **42**, the null hypothesis is **rejected**. As a result, the mean of results from the 5-NN.PO model with bibliography-augmented CN3-term full-text configuration was better than the mean from the SVD CBF baseline with 200 latent topics, statistically different at a **p-value < 0.01** level of significance.

6.3.3.2 Content-boosted Collaborative Filtering

(a) Bibliography-Augmented CBCF

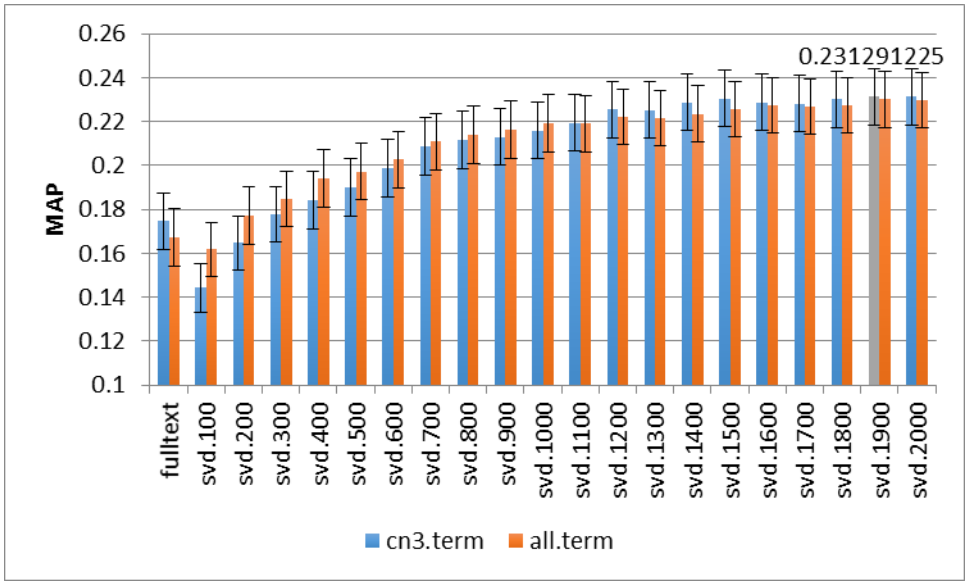


Figure 43: MAP Results of Bibliography CBCF Models

With two TF-IDF vectors for bibliography source, there are four sub types of user-based nearest neighboring content-boosted collaborative filtering models: unigram CN3-term vector space CBCF, unigram All-term vector space, CN3-term latent semantic (SVD), and All-term latent semantic (SVD) CBCF.

As mentioned in the CBF SVD models, the SVD CBCF models were explored to find the optimum *k* latent topics that generated the max mean average precision (MAP) result as depicted in Figure **43**. The MAP results from the CN3-term SVD CBCF bibliography models started with a low MAP, getting better slowly, later staying flat. The MAP result for CN3-term SVD models peaked at the model with 1900 latent topics. The All-term SVD models showed the same pattern

that starting from the low MAP, going up slowly, and then staying flat like on a plateau. The MAP result for All-term SVD models peaked at the model with 1900 latent topics as well.

The CBCF performance between the unigram vector space models showed the CN3-term CBCF slightly better than the All-term unigram CBCF. With respect to the latent semantic models, the CN3-term SVD CBCF model produced MAP results almost the same as the All-term SVD CBCF did. Finally, the maximum MAP was obtained from the CN3-term SVD CBCF model with 1900 latent topics and was therefore selected as the CBCF bibliography model.

(b) Baseline CBCF

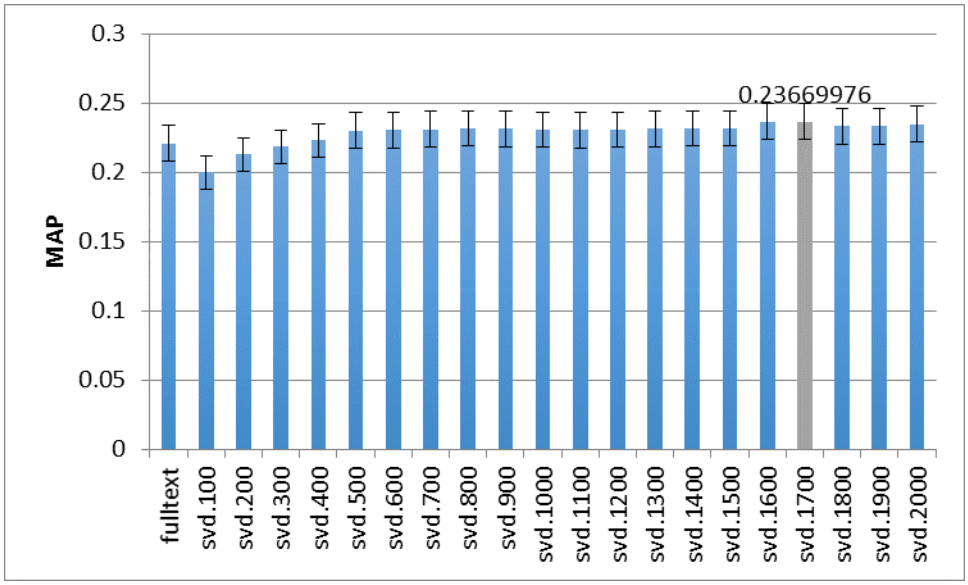


Figure 44: CBF Baselines with Bibliography Users

The CBCF baseline models with 249 users, for whom we retrieved their publications, were assessed as depicted in Figure **44**. Their MAP results stayed steady, and the maximum MAP of the CBCF baseline models came to the CBCF SVD one with 1700 latent topics.

(c) Hypothesis Testing

The individual bibliography-augmented CN3-term SVD CBCF model with 1900 latent topics was assessed with 249 users for whom we retrieved their publications against the SVD CBCF baseline with 1700 latent topics.

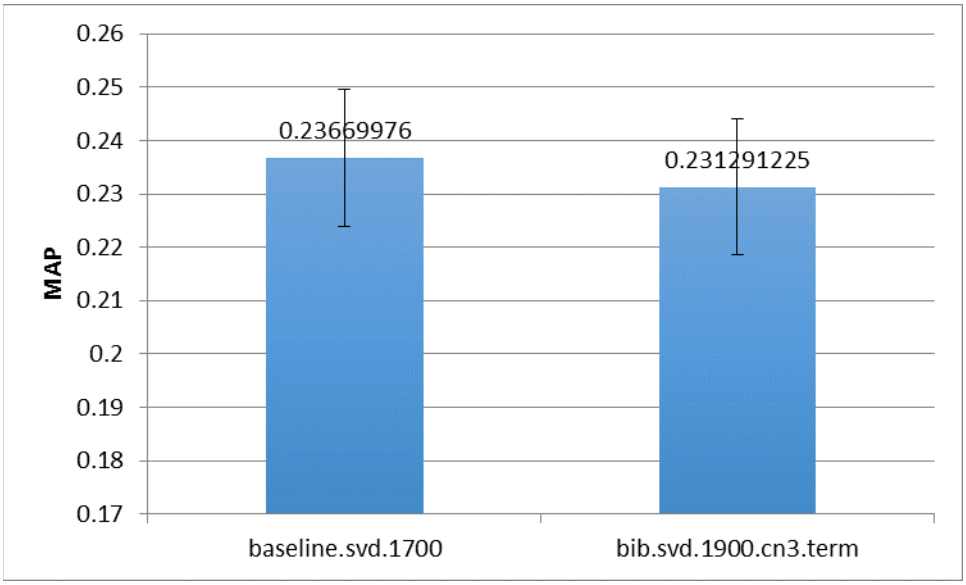


Figure 45: MAP Comparison between Bibliography-Augmented Model and Baseline

H_0: There is no statistical difference between the *accuracy* means of recommended talks from CBCF, *with or without User Publication augmentation*.

One-way Analysis of Variance (ANOVA) was applied to test the hypothesis. The following figure shows the results.

Analysis of Variance (One-Way)

Summary

Groups	*Sample size*	*Sum*	*Mean*	*Variance*
baseline.CBCF.MAP	1245	294.6912	0.2367	0.05388
bib.CBCF.MAP	1245	287.95757	0.23129	0.05296

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*p-level*	*F crit*
Between Groups	0.01821	1	0.01821	0.34086	0.55939	5.41887
Within Groups	132.91401	2488	0.05342			
Total	132.93221	2489				

Figure 46: Hypothesis Testing on Bibliography CBCF Model

From Figure **46**, the null hypothesis is **accepted**. Also, the mean of the results from the bibliography-augmented CN3-term SVD CBCF model with 1900 latent topics is a bit lower than the mean from the SVD CBCF baseline with 1700 latent topics but not statistically different at a p-value < 0.05 significant level.

6.3.4 External Source: External Bookmarks

The experimental models with external scholarly bookmarks augmentation for CBF and CBCF were assessed with 45 users, who provided their external scholarly bookmark accounts with CN3.

6.3.4.1 Content-based Filtering

(a) External Bookmark-Augmented CBF

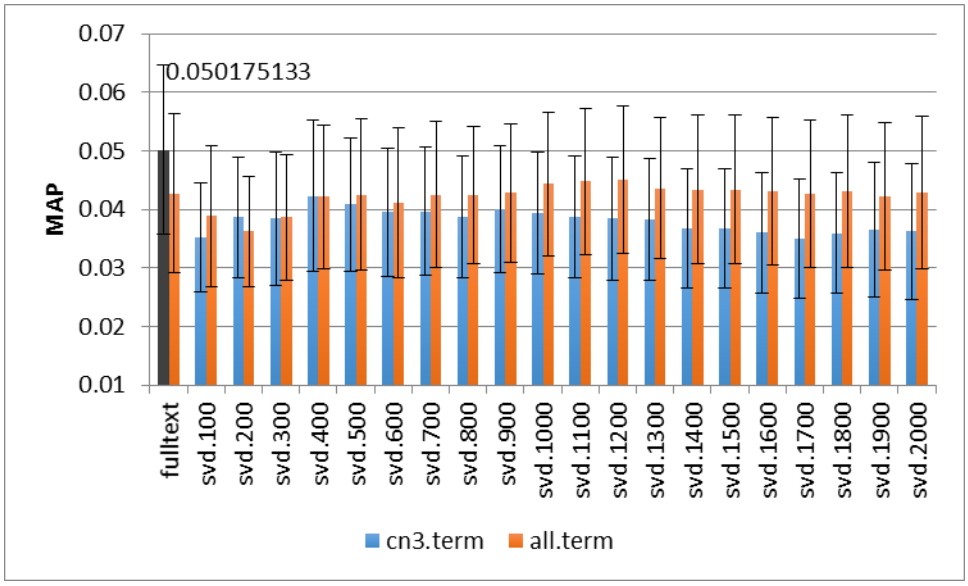


Figure 47: MAP Results of External Bookmark Centroid Models

Similar to the two other external sources, the external bookmark source augmentation models have two TF-IDF vectors, one having only CN3 terms, and another one having all terms from CN3 and extra ones from the external sources. There are three main types of external bookmark-augmented CBF models: individual centroid vector space, clustering centroid, and KNN.PO models. The individual unigram and SVD centroid CBF models on both TF-IDF vectors were explored to find the model that generated the maximum mean average precision (MAP) result as depicted in Figure **47**. The results of the external bookmark-augmented centroid SVD models were the MAP values staying steady and having a peak MAP result at the individual CN3-term unigram model. The numbers of latent topics from the individual CN3-term SVD model at 400

topics and the individual All-term SVD model at 1200 topics were assigned to the other external bookmarked-augmented CBF models as a means of reducing the complexity of the study. Also, the centroid CN3-term unigram centroid model was selected as the representative of the external bookmark-augmented centroid models.

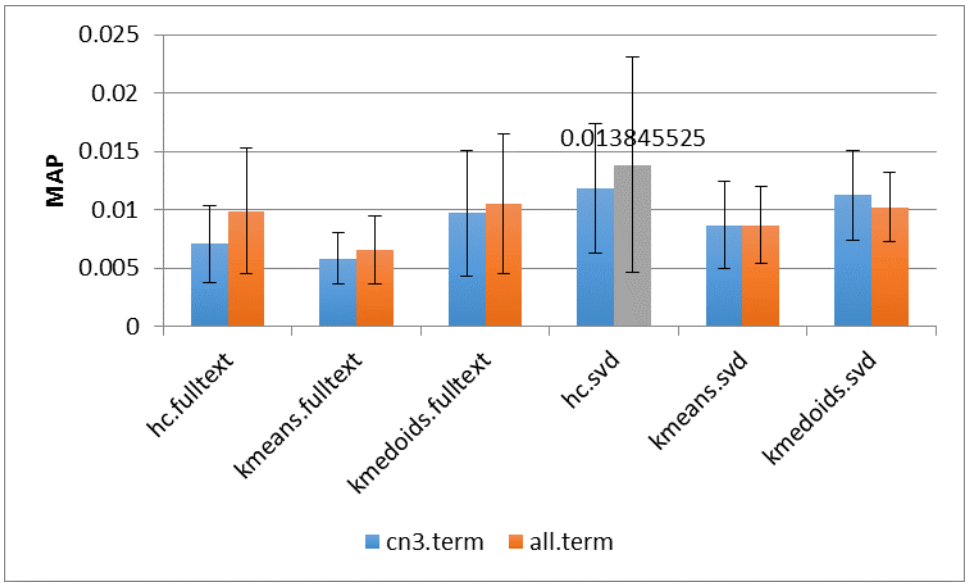


Figure 48: MAP Results of External Bookmark Cluster Models

For the clustering unigram models, as shown in Figure 48, for the CN3-term and the all-term TF-IDF vectors, the hierarchical clustering, K-Means, and K-Medoids model produced poor results. The cluster SVD models used the 400 and 1200 latent topics that were determined by the individual CN3-term SVD and All-term SVD models, respectively. The results from both the clustering SVD centroid models still performed poorly. The external bookmark-augmented hierarchical clustering SVD centroid models were chosen as the representative of the external bookmark-augmented clustering CBF models.

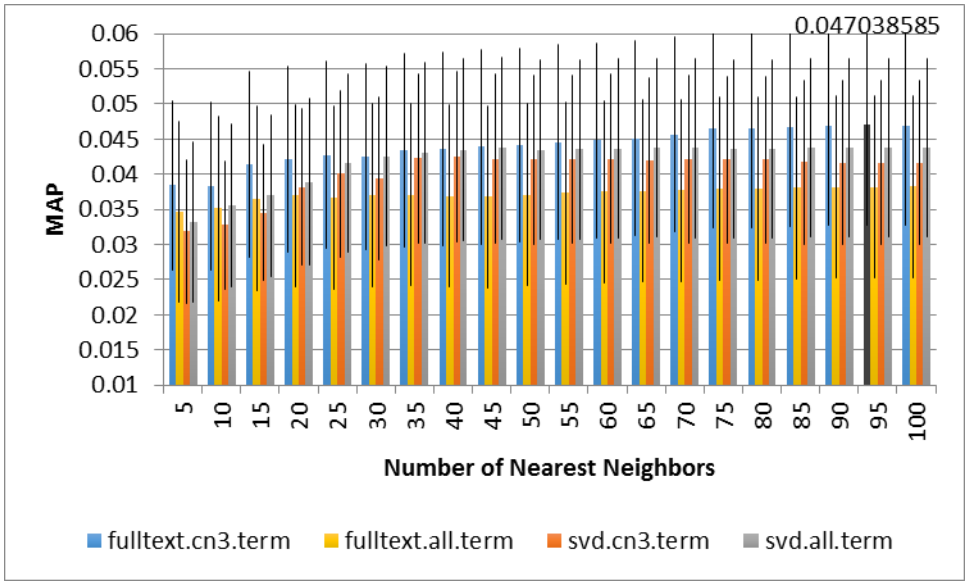


Figure 49: MAP Results of External Bookmark KNN.PO

The MAP results of external bookmark-augmented KNN.PO models using the full-text and latent topics vector space models with CN3-term and All-term vectors were shown in Figure **49**. The results of external bookmark-augmented KNN.PO models with CN3-term vectors performed slightly better than external bookmark-augmented KNN.PO models with All-term vectors on both full-text and SVD ones. The maximum MAP result of external bookmark-augmented KNN.PO models was at the 95-NN.PO CN3-term unigram model. The maximum result was 0.047. As a result, the 95-NN.PO CN3-term unigram model was chosen as the representative of the external-bookmark-augmented KNN.PO models.

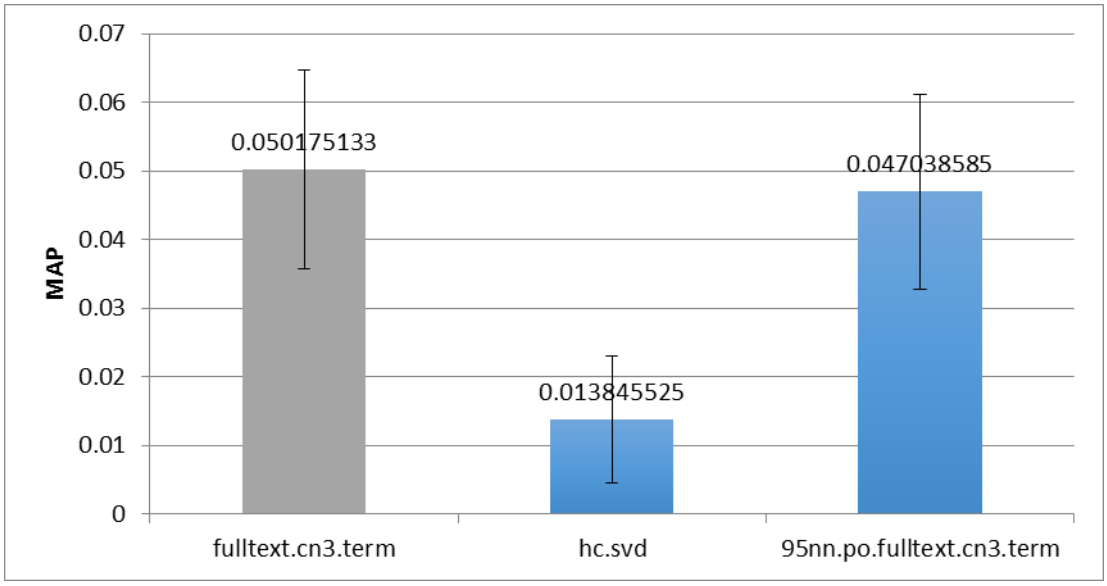


Figure 50: MAP Results of External-Bookmark-Augmented CBF Models

The final external bookmark-augmented CBF comparisons in Figure **50** showed that the individual CN3-term full-text centroid model and 95-NN.PO CN3-term full-text models performed at about the same level of MAP results and outperformed the clustering centroid model significantly. On the other hand, the hierarchical clustering centroid models produced results that were worse than all individual models. In this step, the individual CN3-term full-text vector space model was chosen as the final External Bookmark CBF model.

(b) Baseline CBF

The CBF baselines with the same 45 users, who provided their external bookmark accounts, were assessed as shown in Figure **51**. The maximum MAP result of individual centroid models was 0.0588 on the unigram centroid model, whose MAP result was the peak compared to all the SVD centroid model as shown in the top left diagram on Figure **51**. Also, the maximum MAP result on the individual SVD centroid models was at the model with 300 latent topics. This

number of topics was used later in the other SVD models. On the top right of Figure **51**, all the three SVD clustering models performed better than the unigram clustering centroid models. The MAP result of K-Medoids SVD model was highest among the clustering centroid models. On the bottom left of Figure **51**, all the KNN.PO models performed similarly with no significant differences among them. The maximum MAP result of KNN.PO model was 0.0554 on the 15-NN.PO SVD model.

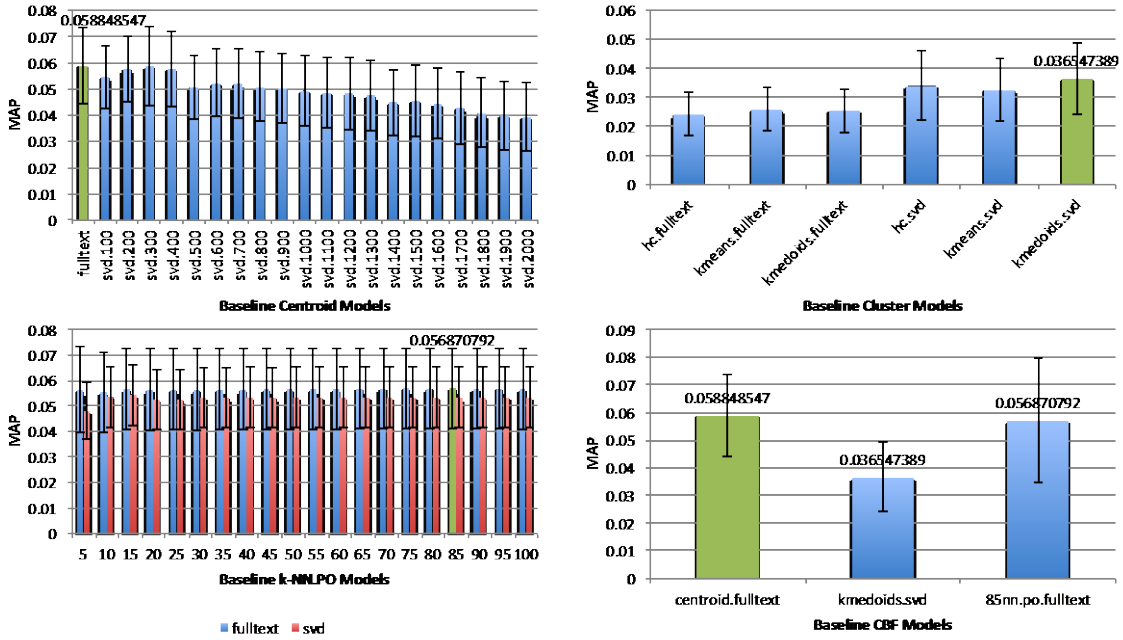


Figure 51: CBF Baselines with External Bookmark Users

The final comparison on the bottom right of Figure **51** showed that the individual unigram centroid model and the 15-NN.PO SVD model performed better than the K-Means SVD model but the result of centroid SVD model was a bit higher than the KNN.PO one. As a result, the individual unigram centroid model was chosen to test the hypothesis.

(c) Hypothesis Testing

The external-bookmark-augmented CN3-term unigram centroid model was assessed with 45 users who provided their external scholarly bookmark accounts with CN3 against the unigram vector space CBF baseline.

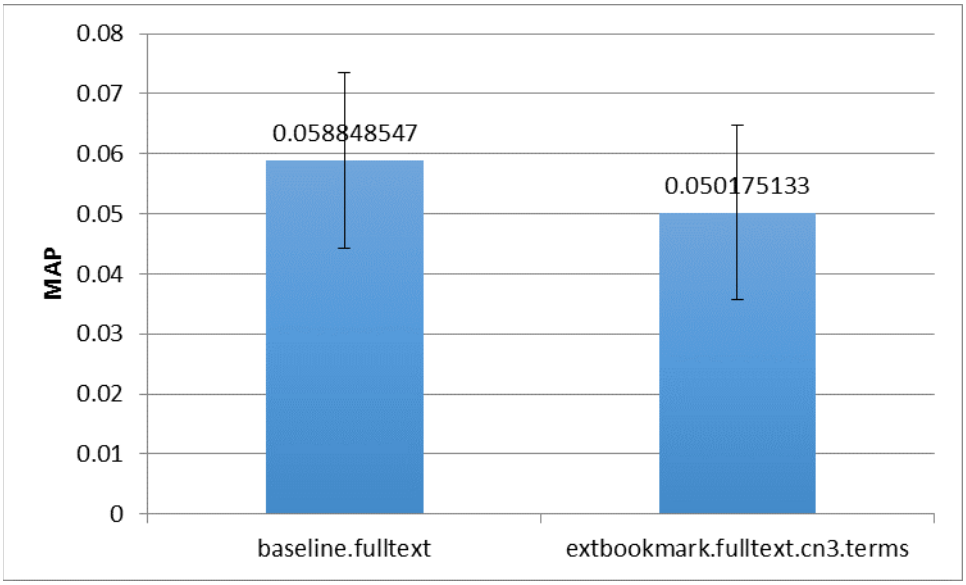


Figure 52: MAP Results of External-Bookmark-Augmented Model vs. Baseline

H_0: There is no statistical difference between the *accuracy* means of recommended talks from CBF, *with or without Bookmarked Scholarly Papers augmentation*.

One-way Analysis of Variance (ANOVA) was applied to test the hypothesis. The following figure shows the results.

Analysis of Variance (One-Way)

Summary

Groups	*Sample size*	*Sum*	*Mean*	*Variance*
baseline.CBF.MAP	225	13.24092	0.05885	0.01257
extbookmark.CBF.MAP	225	11.2894	0.05018	0.0123

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*p-level*	*F crit*
Between Groups	0.00846	1	0.00846	0.68069	0.40979	5.45083
Within Groups	5.57006	448	0.01243			
Total	5.57853	449				

Figure 53: Hypothesis Testing Result on External Bookmark CBF Model

From Figure **53**, the null hypothesis is **accepted**. Also, the mean of result from the external-bookmark-augmented CN3-term unigram vector space CBF model was slightly lower than the mean from the unigram vector space CBF baseline but not statistically different at a p-value < 0.05 significant level.

6.3.4.2 Content-boosted Collaborative Filtering

(a) External Bookmark-Augmented CBCF

With two TF-IDF vectors for external bookmark source, there are four sub types of user-based nearest neighboring content-boosted collaborative filtering models: unigram CN3-term vector space CBCF, unigram All-term vector space, CN3-term latent semantic (SVD), and All-term latent semantic (SVD) CBCF. As mentioned in the CBF SVD models, the SVD CBCF models were explored to find the optimum *k* latent topics that generated the maximum mean average precision (MAP) result as depicted in Figure **54.** The MAP results from the CN3-term SVD CBCF

baseline models started with a low MAP, going up slowly, later staying flat. The MAP result for CN3-term SVD models peaked at the model with 2000 latent topics. The All-term SVD models showed the same pattern of starting from the low MAP, getting better, and then staying steady like on a plateau. The MAP result for All-term SVD models peaked at the model with 1900 latent topics.

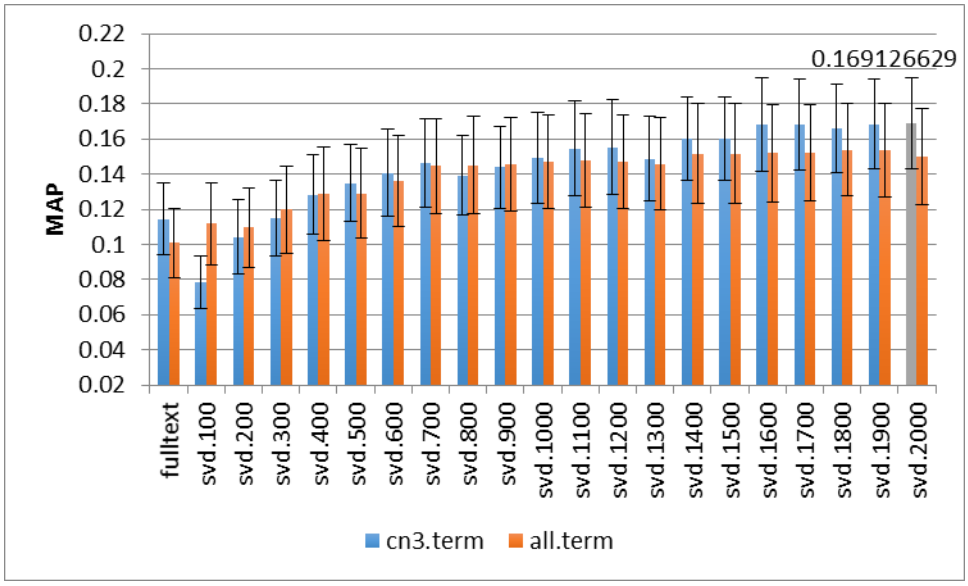


Figure 54: MAP Results of the External Bookmark CBCF Models

The CBCF performance between the unigram vector space models showed the CN3-term CBCF slightly better than the All-term unigram CBCF. Similarly, in the latent semantic models, the MAP result of CN3-term SVD CBCF model was slightly better than one from the All-term SVD CBCF. The maximum MAP was for the CN3-term SVD CBCF model with 2000 latent topics and was selected as the CBCF external bookmark model.

(b) Baseline CBCF

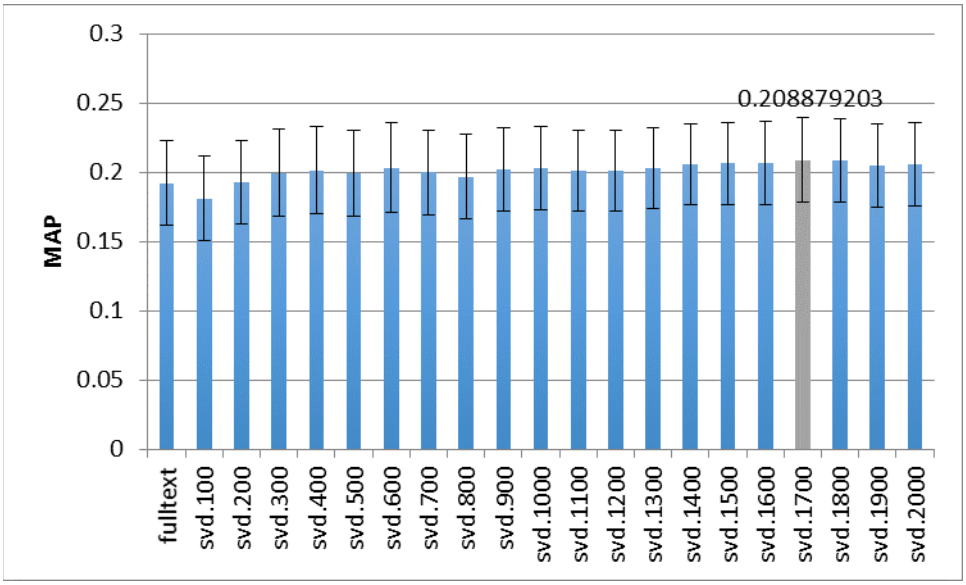


Figure 55: CBCF Baselines with External Bookmark Users

The CBCF baseline models with 45 users, who provided their external bookmark accounts or had their accounts identified by us, were assessed as depicted in Figure **55**. The MAP results were gradually increasing with the number of topics. The maximum MAP of CBCF baseline models was found in the CBCF SVD model, with 1700 latent topics.

(c) Hypothesis Testing

The external-bookmark-augmented CN3-term SVD model with 2000 latent topics was assessed with 45 users who provided their external scholarly bookmark accounts with CN3 against the SVD CBCF baseline with 1700 latent topics.

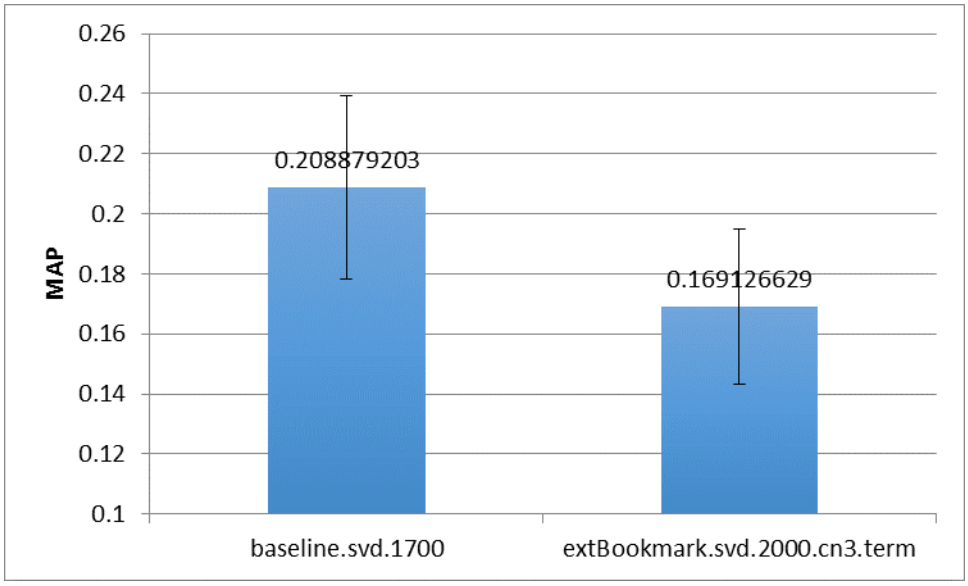


Figure 56: MAP Results of External-Bookmark-Augmented CBCF vs. Baseline

H_0: There is no statistical difference between the *accuracy* means of recommended talks from CBCF, *with or without Bookmarked Scholarly Papers augmentation*.

One-way Analysis of Variance (ANOVA) was applied to test the hypothesis. The following figure shows the results.

Analysis of Variance (One-Way)

Summary

Groups	*Sample size*	*Sum*	*Mean*	*Variance*
baseline.CBCF.MAP	225	46.99782	0.20888	0.05444
extbookmark.CBCF.MAP	225	38.05349	0.16913	0.03912

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*p-level*	*F crit*
Between Groups	0.17778	1	0.17778	3.8007	0.05185	5.45083
Within Groups	20.95547	448	0.04678			
Total	21.13325	449				

Figure 57: Hypothesis Testing of External Bookmark CBCF Model

From Figure **57**, the null hypothesis is **accepted**. Also, the mean of the results from the external-bookmark-augmented CN3-term SVD CBCF model with 2000 latent topics was lower than the mean from the SVD CBCF baseline with 1700 latent topics and almost statistically different at a p-value < 0.05 significant level.

6.4 SUMMARY AND DISCUSSION

Given a small set of users in each conference hosted by CN3, “under-contribution” phenomena in the small communities, and the “Short-life Time-Span Talks” property, recommending research talks in this context is a real challenge to overcome. One of the simple ways to address the challenge is by augmenting the user profiles with external sources. We explored six augmented experimental models (3 CBF + 3 CBCF) comparing them with six baselines (3CBF + 3CBCF).

For homepage augmentation, both CBF and CBCF homepage representative models performed a bit higher than baseline models but there was no significance. In general, adding information from homepage sources did not harm nor help to improve the accuracy performance of recommending models in the research.

In the bibliography source, bibliography-augmented models provided better results. The CBF bibliography representative model, bibliography-augmented 5-NN.PO CN3-term full-text, outperformed the CBF baseline, SVD CBF centroid baseline with 200 latent topics, with p-value less than 0.01. On the other hand, the CBCF bibliography representative model slightly performed a bit lower than the CBCF baseline.

The last external source in this study, both CBF and CBCF representative models with external scholarly bookmarked papers augmentation yielded poor results. Both of them performed a bit lower than baseline models.

When looking at each accuracy performance in CBF recommending approach, no cluster model was selected as a representative model for any external source augmentation or baseline ones. Apparently cluster model in all three clustering methods with the space dimension (full-text vs. latent space) or feature selection (considering All terms from CN3 and extra terms introduced from external sources or excluding only target terms from CN3 corpus) underperformed other external-bookmark-augmented centroid or KNN.PO models.

Also, almost all of the representative CBF models were the centroid models. Only the bibliography-augmented 5-NN.PO models with CN3-term full-text configuration was selected as the representative of the bibliography CBF models that were not a centroid model.

In the CBCF models, models with or without external source augmentations performed with no significant differences to one another. There was one external-bookmark-augmented

CBCF model that performed slightly lower than the CBCF baseline one with almost statistical significance (p-value = 0.052).

In the latent semantic settings, there were only three representative models (two experimental CBF models and one CBF baseline one) that came with full-text setting. Other nine representative models were ones with latent space user profile representation.

In the unigram terms (features) selection setting, five out of six experimental models (3 CBF + 2 CBCF) with the CN3-term settings (considering only terms appearing in the CN3 corpus). The research shows that excluding extra terms introduced from the external sources help improve the accuracy performance.

Discussion

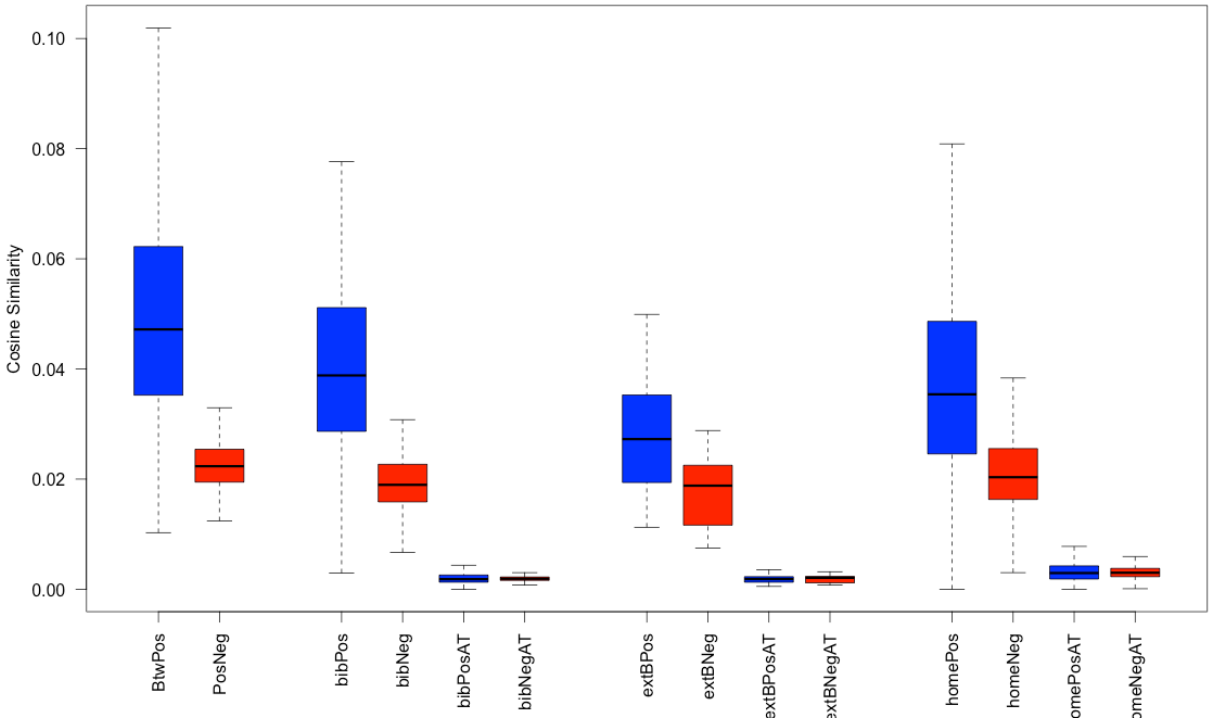


Figure 58: Full-Text Similarities between Bookmarked CN3 Talks and External Sources

1) The first blue boxplot on the left is the average similarity of each user between their bookmarked CN3 talks themselves.
2) The first red boxplot on the left is the average similarity of each user between bookmarked CN3 talks and non-bookmarked ones.
3) The second blue boxplot on the left is the average similarity of each bibliography-equipped user between their bibliography and the bookmarked CN3 talks on the excluded CN3-term only.
4) The second red boxplot on the left is the average similarity of each bibliography-equipped user between their bibliography and the non-bookmarked CN3 talks on the excluded CN3-term only.
5) The third blue boxplot on the left is the average similarity of each bibliography-equipped user between their bibliography and the bookmarked CN3 talks on the All-term, including extra ones introduced from the bibliography corpus.
6) The third red boxplot on the left is the average similarity of each bibliography-equipped user between their bibliography and the non-bookmarked CN3 talks on the All-term, including extra ones introduced from the bibliography corpus.
7) The fourth blue boxplot on the left is the average similarity of each external-bookmark-equipped user between their external bookmarked papers and the bookmarked CN3 talks on the excluded CN3-term only.
8) The fourth red boxplot on the left is the average similarity of each external-bookmark-equipped user between their external bookmarked papers and the non-bookmarked CN3 talks on the excluded CN3-term only.

9) The fifth blue boxplot on the left is the average similarity of each external-bookmark-equipped user between their external bookmarked papers and the bookmarked CN3 talks on the All-term, including extra ones introduced from the external bookmarked papers corpus.
10) The fifth red boxplot on the left is the average similarity of each external-bookmark-equipped user between their external bookmarked papers and the non-bookmarked CN3 talks on the All-term, including extra ones introduced from the external bookmarked papers corpus.
11) The sixth blue boxplot on the left is the average similarity of each homepage-equipped user between their homepage and the bookmarked CN3 talks on the excluded CN3-term only.
12) The sixth red boxplot on the left is the average similarity of each homepage-equipped user between their homepage and the non-bookmarked CN3 talks on the excluded CN3-term only.
13) The seventh blue boxplot on the left is the average similarity of each homepage-equipped user between their homepage and the bookmarked CN3 talks on the All-term, including extra ones introduced from the homepage corpus.
14) The seventh red boxplot on the left is the average similarity of each homepage-equipped user between their homepage and the non-bookmarked CN3 talks on the All-term, including extra ones introduced from the homepage corpus.

For content-based recommendation with external source augmentation (CBF), the bibliography-augmented 5-NN.PO CN3-term full-text approach was the only one model that beat

the CBF baseline with p-value less than 0.01. As shown in the third and fourth boxplots from the left in Figure **58**, the CN3-term full-text user publications were closer to the bookmarked talks and further away from the un-bookmarked talks. The bibliographies were even a bit farther from the un-bookmarked talks (cosine similarity: 0.0192) than the bookmarked talks were (cosine similarity: 0.0224). One hypothesis was the conference attendees came to the venue with a purpose to attend the talks related to their research interest. From the results, all three sources showed that they contained user research interests, but the bibliography source provided the best one out of them. Surprisingly, the homepage-augmented CBF approach was the last one to be expected to exceed the baseline. The representative CBF models with homepage augmentation yielded the same MAP results as the baselines did. The hypothesis was because the homepages partly contained the publications of the users even though the rest of their pages contained other users' interests.

Figure **59** shows the boxplot of the average latent semantic similarity, for each user, between bookmarked CN3 talks, and non-bookmarked talks or between documents of each external source and non-bookmarked CN3 talks from selected models. The boxplots have the same order as Figure **58**.

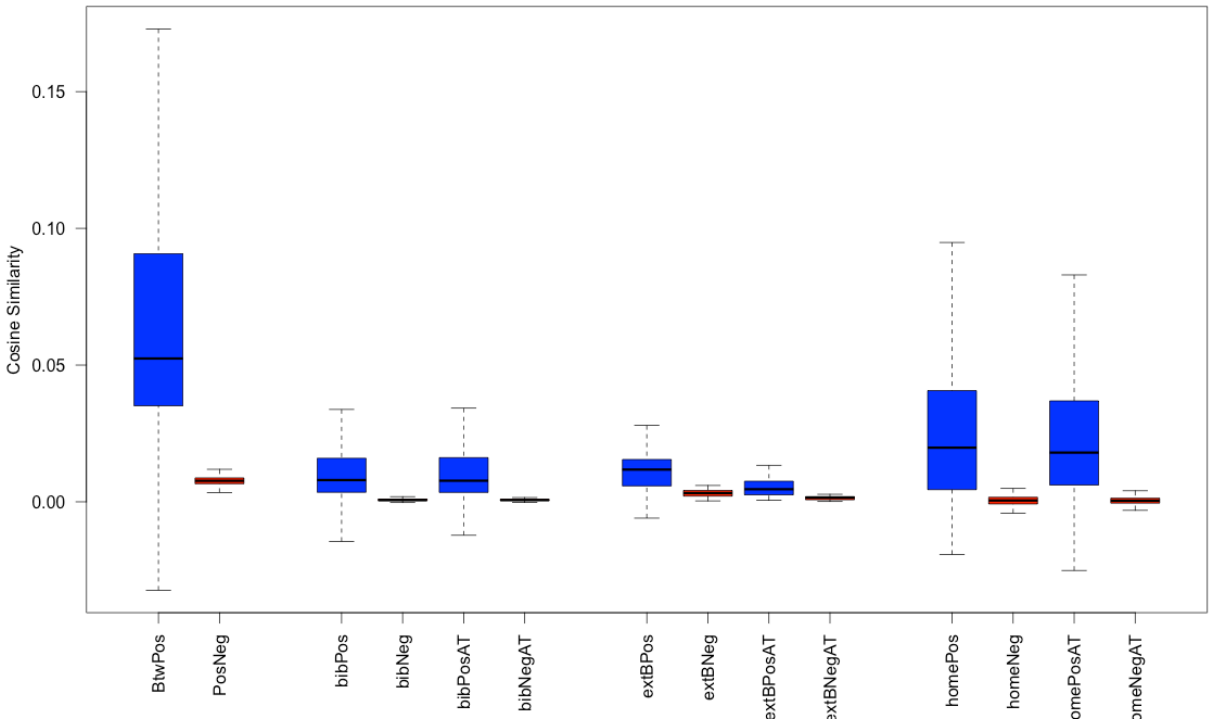


Figure 59: Latent Semantic Similarities between Bookmarked CN3 Talks and External Sources

As shown in the eleventh and twelfth boxplots from the left in Figure **59**, the CN3-term SVD user homepages were closer to the bookmarked talks and also far from the un-bookmarked talks. On the other hand, the external scholarly papers reflected more about the general user interests. From the preliminary result, the external scholarly papers played a role to improve the recommendation performance because the different nature of attending user behaviors and time constraints. From Figure **58**, the external scholarly papers were not a good indicator to distinguish between bookmarked talks and un-bookmarked talks as their similarities depicted in the seventh and eighth boxplots were closer to each other than the bookmarked talks against the un-bookmarked one shown in the first and second boxplot, respectively. From the results, one of the

hypotheses that can be made was that in conference attending situations, attendees are more likely to prefer attending talks more related to their research interests. While attending academic talks was another story, users would go to the talks if they had spare time and more likely to join the talks that share their general interests. Another hypothesis about the poor performance of external-bookmark models was the introduction of noise or terms that were related to the target corpus. In general, it is easy to bookmark interesting papers into the bookmarking systems, whether those papers are very close to the research interest or just related to their general interests. On the other hand, publishing papers is much harder than bookmarking them, also these bibliographies concentrate with terms or keywords related to their research interests.

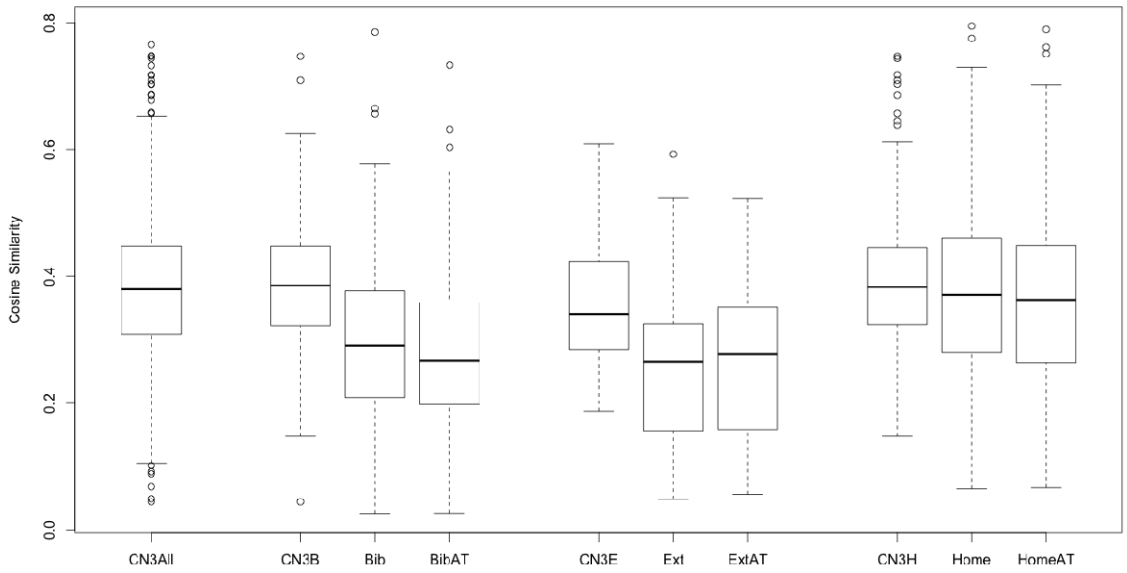


Figure 60: Latent Semantic Similarities between User Profiles and Their 10 Nearest Peers

Figure **60** shows the boxplot of the average latent semantic similarity, between user profiles and their ten nearest peers in CBCF models with and without external source augmentations.

1) The first boxplot on the left is the average similarity between user profiles of CBCF baseline model and their ten nearest peers.
2) The second boxplot on the left is the average similarity between user profiles of CBCF baseline model with bibliography users and their ten nearest peers.
3) The third boxplot on the left is the average similarity between user profiles of CBCF bibliography-augmented CN3-term model with bibliography users and their ten nearest peers.
4) The fourth boxplot on the left is the average similarity between user profiles of CBCF bibliography-augmented All-term model with bibliography users and their ten nearest peers.
5) The fifth boxplot on the left is the average similarity between user profiles of CBCF baseline model with external-bookmark users and their ten nearest peers.
6) The sixth boxplot on the left is the average similarity between user profiles of CBCF external-bookmark-augmented CN3-term model with external-bookmark users and their ten nearest peers.
7) The seventh boxplot on the left is the average similarity between user profiles of CBCF external-bookmark-augmented All-term model with external-bookmark users and their ten nearest peers.
8) The eighth boxplot on the left is the average similarity between user profiles of CBCF baseline model with homepage users and their ten nearest peers.
9) The ninth boxplot on the left is the average similarity between user profiles of CBCF homepage-augmented CN3-term model with homepage users and their ten nearest peers.

10) The tenth boxplot on the left is the average similarity between user profiles of CBCF homepage-augmented All-term model with homepage users and their ten nearest peers.

For the content-boosted collaborative filtering (CBCF), the CBCF approach without any augmentation performed relatively higher than ones with external source augmentation thanks to a bit higher similarity between users and their top ten nearest peers (mean cosine similarity: 0.37978) from 815 users from CN3 dataset as shown in the first boxplot in Figure **60** than external-source-augmented ones. The external sources, bibliography and external scholarly papers in particular, did not provide any impact to the user similarity matrix in order to improve the recommendation results in the general situation but reduced the similarity between users and their peers except in the homepage-augmented models. As expected, the SVD CBCF approach outperformed the unigram vector space CBCF models especially in bibliography on both TF-IDF vectors and homepage on All-term TF-IDF vector.

The Bibliography-augmented and the homepage-augmented CN3-term unigram vector space models statistically increased the MAP performance compared with the same-source vector space CBF models with All-term vectors. Also, external-bookmark-augmented CN3-term unigram vector space model produced the means of MAP result higher than other external-bookmark-augmented ones. There are two main factors to explain these performances. One reason is that the term frequency (TF) with extra terms, introduced by external sources, increases the sparseness to the sparseness to the vectors themselves. Another is that the inverse document frequency (IDF) taking the extra documents from the external sources into account dilutes the importance of the “where-about” of the terms in the target CN3 talks. While the CN3-term vector space models produced MAP results that are not statistically different from the same-source

individual SVD CBF models with both TF-IDF vectors, CN3-term vector space models can produce the recommendation list in seconds. Both advantages combined make these models, having the same quality with SVD CBF ones and being faster to produce recommendation because they do not need to conduct the SVD transformation, which takes hours to complete, attractive. This CBF method is more appealing to use given its simple approach and implementation speed to produce the recommendation lists.

Implications

Even though there was only one case where external-source-augmented CBF models produced statistically better recommendations than the CBF baselines did, there are two aspects to note. One is that two out of six experimental models (bibliography-augmented 5-NN.PO CN3-term unigram CBF and homepage-augmented All-term SVD CBCF) achieved higher MAP means than the baselines, which is taken as the input to the next study, the recommendation fusion. Another was that the means of experimental CBF models were about the same as the baseline *when full in-system user profiles were considered*. This means that in the cold-start situations, the experimental CBF with extra user information has a good chance to overcome the baseline.

7.0 STUDY 2: COLD-START PROBLEM

The cold-start problem is one of the common problems that recommender systems face when systems get new users. One of the main factors that affect the quality of recommendations is the amount of user preference information available to the recommender systems. In order to gain more information about users, the external sources come into play as one of the crucial roles. This study aims to explore the impact of the external source augmentation on the recommendation results.

7.1 DATASET DEMOGRAPHICS

From Figure **12**, there were 434 users qualified who provided or we identified at least one external source of theirs and bookmarked at least five talks in any conference, consisting of

1. The personal webpage (317 users),
2. Google Citation for the user publication source (170 users) and Scopus (79 users),
3. CiteULike (28 users) and Mendeley (17 users) accounts are provided in the CN3 user personal information database.

7.2 STUDY PROCEDURE

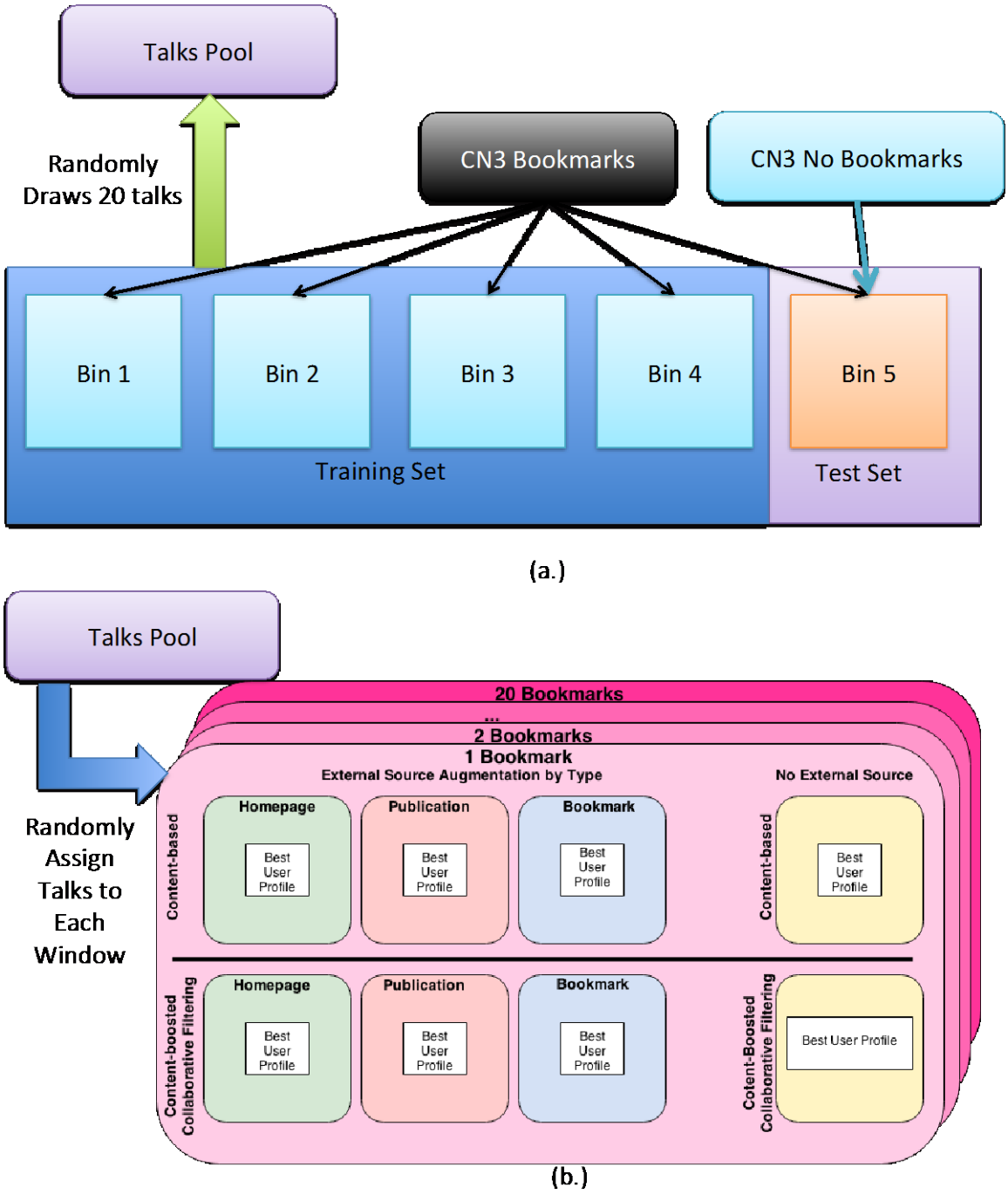


Figure 61: Cold-Start Effect Study

Objective. The purposes for injecting external sources into recommender systems is to increase the knowledge regarding to their users in order to provide a way to improve the recommendation performance. This experiment is intended to understand the effect of the cold-start issue after recommender systems have been enhanced with external source augmentations.

Setting. In order to study the cold-start effect, we assigned the users into 21 different size bins according to the number of bookmarks in the CN3 system platform, ranging from, no bookmark at all, to twenty bookmarks in the training set. Each user could be assigned to several bins. The numbers of bins users were assigned to, depended on the bookmarks they have in their user profiles. For example, users who have twenty bookmarks might have up to sixteen bookmarks for training and four bookmarks for testing, for which they were assigned to the first seventeen bins.

Figure **61** shows the diagram of the cold-start effect simulation study. With respect to experimental settings, there are three conditions in this study: Cold-start

1. By given the type of external sources (homepage, user publications or external bookmarks),
2. By given the size of bookmark bin (to simulate how sparse the user profile is),
3. By given the type of recommender systems (CBF vs. CBCF).

For each user, as seen in Figure **61**a, the bookmarked talks were split into five folds randomly. Each fold in turn was used for recommendation assessment along with un-bookmarked talks. The remaining four folds were used to construct the experimental recommendation approaches. If the total bookmarked talks were more than twenty, they were randomly selected only twenty ones.

Secondly, as shown in Figure **61**b, randomly picked talks in the training set according to the cold-start situation. The random selection was repeated ten times for every bin. The intermediate result for each bin was the average of these ten iterations. The evaluation result was the average of five testing assessments.

The recommending approaches were brought from the study 1 as below:

1. **Homepage**

Experimental CBF: Homepage-augmented 20-NN.PO SVD CN3-term CBF model

Baseline CBF: Centroid SVD CBF model with 100 topics

Experimental CBCF: Homepage-augmented SVD CBCF model with 900 topics

Baseline CBCF: SVD CBCF model with 1600 topics

2. **User Publication (Bibliography)**

Experimental CBF: Bibliography-augmented 5-NN.PO Full-text CN3-term CBF model

Baseline CBF: Centroid SVD model with 200 topics

Experimental CBCF: Bibliography-augmented SVD CBCF model with 1900 topics

Baseline CBCF: SVD CBCF model with 1700 topics

3. **Bookmarked Scholarly Papers (External Bookmark)**

Experimental CBF: External Bookmark-augmented Centroid Full-text CN3-term Model

Baseline CBF: Centroid Full-text model

Experimental CBCF: External Bookmark-augmented SVD CN3-term CBCF model with 2000 topics

Baseline CBCF: SVD CBCF model with 1700 topics

Dependent Variables. The dependent variable is the MAP. The recommendations returned from 126 experimental approaches (6 approaches x 21 bins) with external sources are expected to provide higher MAP than the recommendation results returned from 126 no-external-source baselines (6 baselines x 21 bins).

Hypotheses of Study 2

Research Question 2

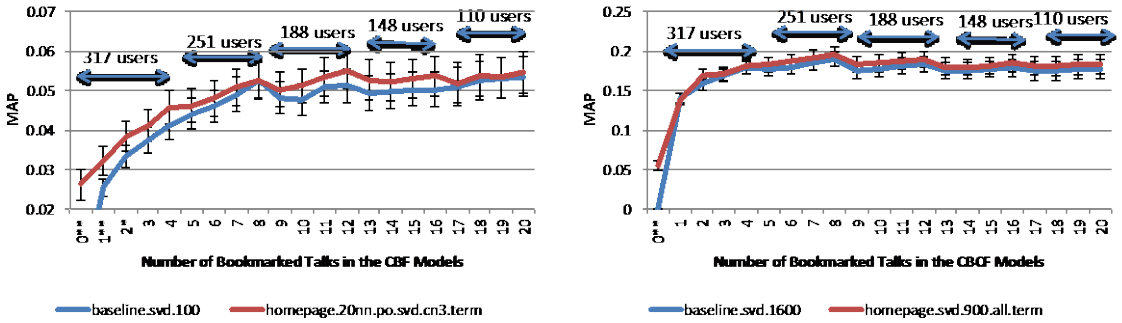
"*Could the external sources help alleviate the cold start situation in the research talk recommendation?*"

H_0: There is no statistical difference between the *accuracy* means of recommended talks from any CBF or CBCF, with *different size of bin*, and,

	CBF	**CBCF**
Homepage	*with or without Personal Webpage augmentation.*	
User Publication	*with or without User Publication augmentation*	
Bookmarked Scholarly Papers	*with or without Bookmarked Scholarly papers augmentation*	

7.3 RESULTS

7.3.1 Homepage



* Indicates significant level p-value < 0.05; ** indicates significant level p-value < 0.01

Figure 62: Homepage-Augmented Recommendations on the Cold-Start Effect

The homepage-augmented 20-NN.PO SVD CN3-term CBF model and the centroid SVD baseline CBF model with 100 latent topics were carried from the study 1 in order to study the cold-start problem. So did the homepage-augmented SVD All-term CBCF model and the SVD baseline CBCF one with 1600 latent topics. These four models were assessed with 317 CN3 users who provided their homepage information or we identified theirs. Recommendations with homepage augmentation on both CBF and CBCF performed significantly better than the baselines in the very start of cold-start situations when users have no bookmarks. Homepage-augmented CBF also outperformed in other two situations that users having one and two bookmarks with p-value = 0.003 and 0.03449, respectively. After that, both of them performed slightly better than baselines but not statistically different.

7.3.2 User Publication (Bibliography)

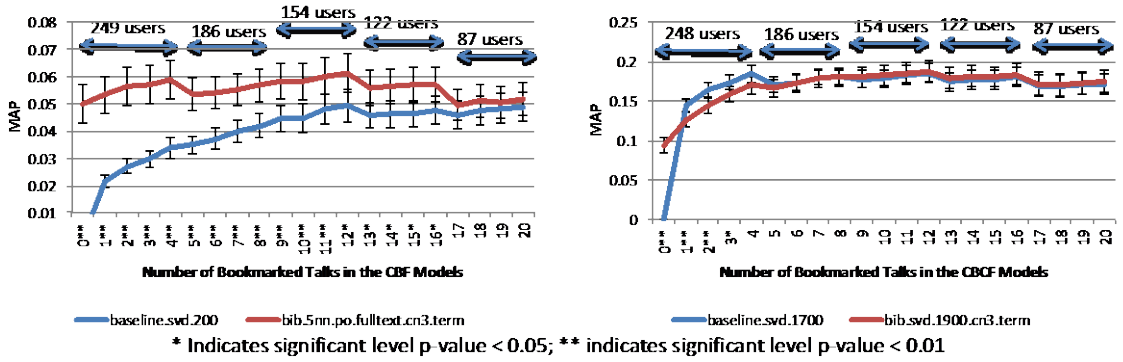


Figure 63: Bibliography-Augmented Recommendations on the Cold-Start Effect

The bibliography-augmented 5-NN.PO full-text CN3-term CBF model and the centroid SVD baseline CBF model with 200 latent topics were carried from the study 1 in order to study the cold-start problem. So did the bibliography-augmented SVD CN3-term CBCF model and the SVD baseline CBCF one with 1700 latent topics. These four models were assessed with 249 CN3 users who provided their publications or we identified theirs. Like the homepage evaluation results, bibliography-augmented recommendations on both CBF and CBCF outperformed baselines significantly in the no-bookmark cold-start situation. Moreover, bibliography-augmented CBF kept significantly outperforming the CBF baseline from no bookmark until eleven bookmarked talks with significance level p-values < 0.01 and sixteen bookmarked talks in the user profile with significance level p-values < 0.05. Bibliography-augmented CBCF counterparts performed significantly better than the CBCF baseline one when users have no bookmarked talks. However, bibliography-augmented CBCF ones underperformed significantly

in the early cold-start stage from one to three bookmarks. After that, both performed approximately at the same level.

7.3.3 Bookmarked Scholarly Papers (External Bookmark)

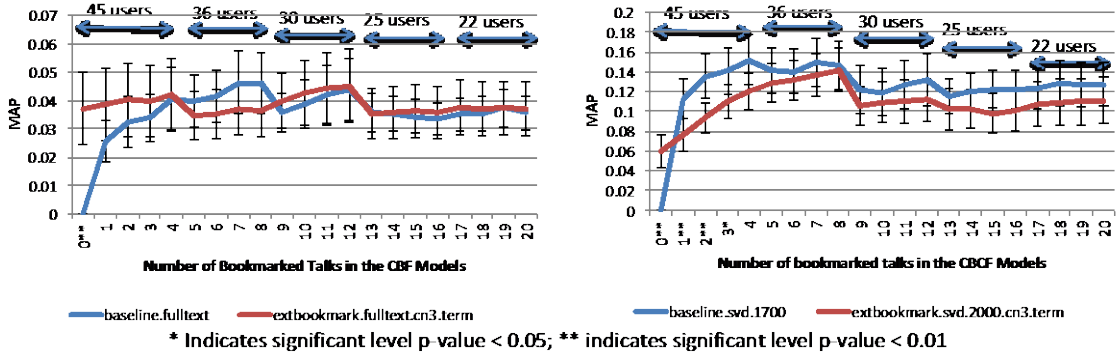


Figure 64: External-Bookmark-Augmented Recommendations on the Cold-Start Effect

The external bookmark-augmented centroid full-text CN3-term CBF model and the centroid full-text baseline CBF model were carried from the study 1 in order to study the cold-start problem. So did the external bookmark-augmented SVD CN3-term CBCF model and the SVD baseline CBCF one with 2000 latent topics. These four models were assessed with 45 CN3 users who provided their external bookmark accounts or we identified theirs. Recommendations with external bookmarked scholarly papers augmentation on both CBF and CBCF improved significantly in the starting cold-start situation that users have no bookmarks. After that, external-bookmark-augmented CBF performed not statistically different from baselines. However, external-bookmark-augmented CBCF performed worse than baselines from one to three bookmarks.

7.4 SUMMARY AND DISCUSSION

This cold-start problem study showed that the external source augmentation helped alleviate the cold-start problem on both content-based and content-boosted collaborative filtering models. Homepage-augmented, bibliography-augmented, and external bookmark-augmented CBF models performed significantly better than the CBF baseline in the early stages of cold-start problems (window size: 0 – 2, 0 – 16, and 0 respectively). After the early stage when external source augmentation CBF models outperformed the baselines, the models still slightly performed better than the baselines.

The external source-augmented CBCF models helped alleviate the cold-start problem only when users had no bookmark. However, after users started having bookmarked talks in the profile, Bibliography-augmented, and external bookmark-augmented CBCF models underperformed the CBCF baselines in the early stages (window size: 1 – 3, and 1– 3, respectively). After that there were no statistical differences in the MAP results on all experimental models comparing with the CBCF baselines.

Implications

These external source augmentation CBF models would benefit for the small-community recommender systems that just launch and has no or few users. With too few users, CBCF models do not work efficiently. However, the systems, that have been established for a certain amount of time and already had active users, could take advantage of the external-source-augmented CBCF models when users have not yet bookmarked in order to boost up the first recommendation to the new users.

8.0 STUDY 3: RECOMMENDATION FUSION

The recommendation fusion models combine strong aspects from different models such as the content-based recommendation with recommending items oriented on the user profiles or the collaborative filtering from the same-minded people. Study 1 - external-source-augmented recommendation improvement showed that some recommendation approaches performed better with user information from external sources. This chapter studies the impact of combining these approaches augmented by different sources and different methods, or the same source with difference methods.

8.1 DATASET DEMOGRAPHICS

From the CN3 dataset that we used for the study 1, 159 users, who were qualified to use in this study, had at least two external sources, as showed in the Figure 65 in bold face. Five-fold cross-validation assessment was conducted in this study. The qualified users had at least five bookmarked talks in any conference and also provided at least two external sources, consisting of: a personal webpage (150 users); user publication sources, including Google Citation (106 users) and Scopus (41 users); and/or external scholarly bookmark sources, including CiteULike

(26 users) and Mendeley (13 users) accounts. These sources were provided in the CN3 user personal information, or we identified their external sources proactively.

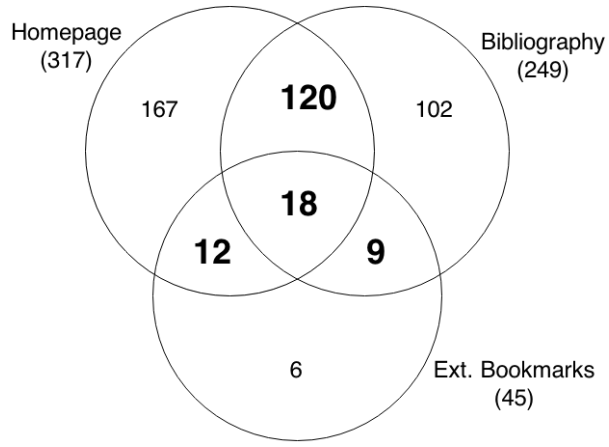


Figure 65: Users in Bold Face Demography in CN3 in the Recommendation Fusion Study

8.2 STUDY PROCEDURE

Objective. To understand the effect of fusing recommendation approaches that use different external sources. The six recommendation approaches that are carried away from Study 1 (Figure **18**) as a source are used in this experiment. These six models of external source augmentation approaches were fused in cross-source and same-source fashions as shown in Figure **66**b and Figure **66**c.

Setting. Six experimental recommendation approaches and six baselines were determined from Study 1 as shown in Figure **18**. There were nine experimental (six different-source-different-approach, and three same-source models) combinations of data fusion models. Two score-based

data fusion methods (*CombSUM* and *CombMNZ*) were used. The weighting step in the data fusion methods was 0.1.

The recommending approaches brought from the study 1 are the following:

1. **Homepage**

Experimental CBF: Homepage-augmented 20-NN.PO SVD CN3-term CBF model

Baseline CBF: Centroid SVD CBF model with 100 topics

Experimental CBCF: Homepage-augmented SVD CBCF model with 900 topics

Baseline CBCF: SVD CBCF model with 1600 topics

2. **User Publication (Bibliography)**

Experimental CBF: Bibliography-augmented 5-NN.PO Full-text CN3-term CBF model

Baseline CBF: Centroid SVD model with 200 topics

Experimental CBCF: Bibliography-augmented SVD CBCF model with 1900 topics

Baseline CBCF: SVD CBCF model with 1700 topics

3. **Bookmarked Scholarly Papers (External Bookmark)**

Experimental CBF: External Bookmark-augmented Centroid Full-text CN3-term Model

Baseline CBF: Centroid Full-text model

Experimental CBCF: External Bookmark-augmented SVD CN3-term CBCF model with 2000 topics

Baseline CBCF: SVD CBCF model with 1700 topics

While the same-source baseline models were determined from Study 1, as mentioned above, the cross-source and the same-source models on CBF and CBCF were reassessed on the subset of users for whom these external sources were available.

Five-fold cross-validation was conducted the same way as in Study 1. The 3321 talks were split into five folds randomly. Each fold was used in turn for recommendation assessment; the remaining four folds were used to construct the experimental recommendation approaches in each experiment. The evaluation result was the average of the results of five testing assessments.

Dependent Variables. The dependent variable is the MAP. The recommendations returned from twelve experimental approaches with external source augmentation are expected to provide higher MAP than the recommendation results returned from two no-external-source baselines and one best fusion baseline.

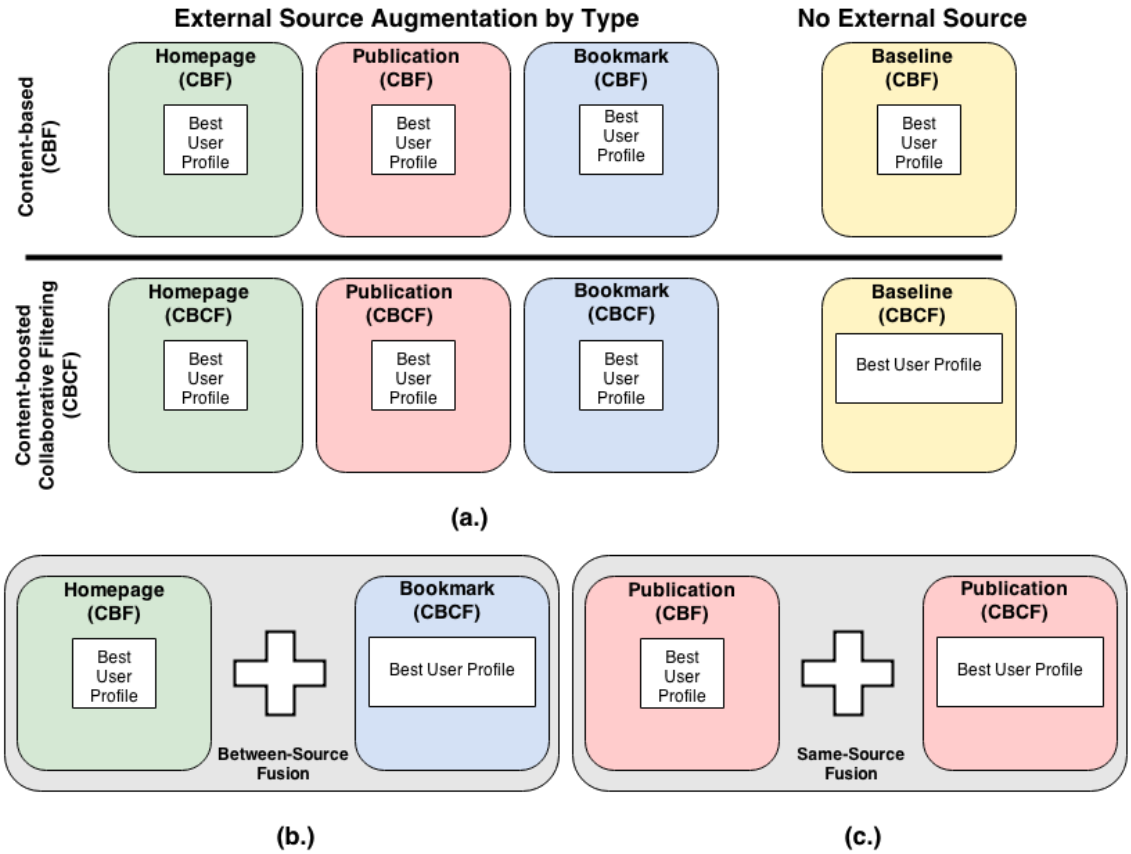


Figure 66: Data Fusion Demonstration

Hypothesis of Study 3

Research Question 3

"*Which combinations of the different recommendation approaches generate better recommendation result?*"

H_0: There is no statistical difference between the *accuracy* means of recommended talks from any fusion recommendation approach with external source augmentation and the baseline approaches.

Metrics: ***MAP***

8.3 RESULTS

8.3.1 Different-Source-Different-Approach Fusion

8.3.1.1 Homepage-Bibliography Fusion

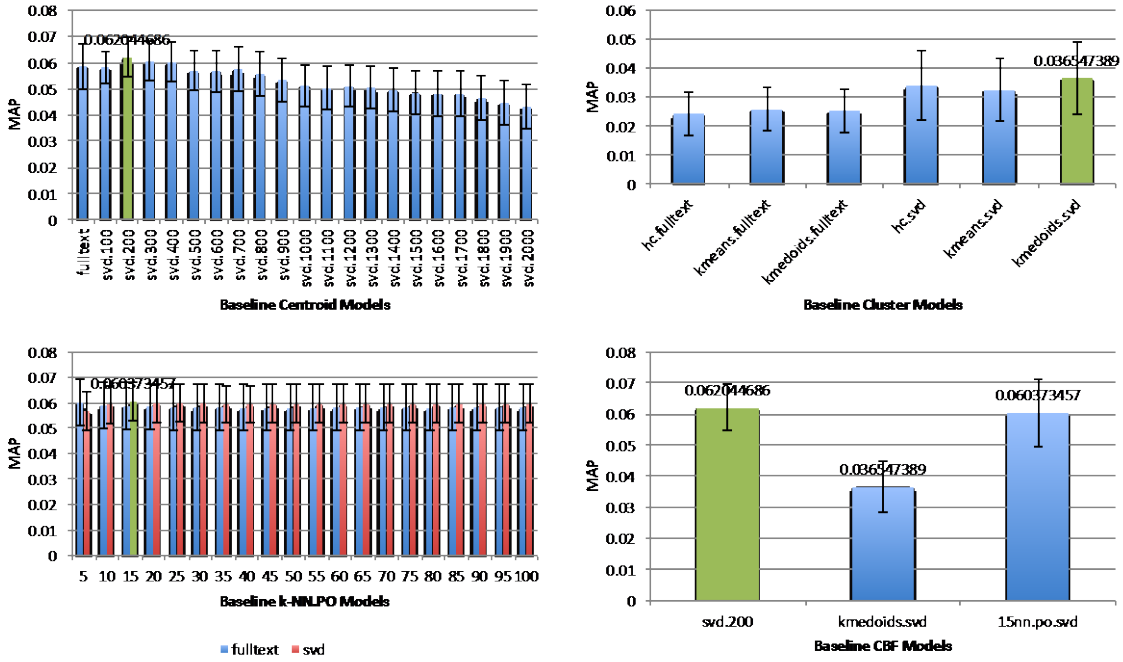


Figure 67: CBF Baselines on 138 Homepage + Bibliography Fusion Users

The CBF baselines with the 138 users, who provided their homepage and publications, or had their homepages and publications identified by us, were assessed as shown in Figure **67**. The maximum MAP result of the individual centroid models was 0.062 on the SVD centroid model with 200 latent topics. The top left diagram in Figure **67** shows the peak MAP result compared to the other SVD centroid and the full-text centroid models. The other SVD CBF baseline models

also used 200 latent topics. As shown in the top right of Figure **67**, all three SVD clustering models performed better than the full-text clustering centroid models. The MAP result of the K-Medoids SVD model was the highest among the clustering centroid models. On the bottom left of Figure **67**, all the KNN.PO baseline models performed similarly with no significant differences among them. The maximum MAP result of the KNN.PO model was 0.0604 on the 15-NN.PO SVD model.

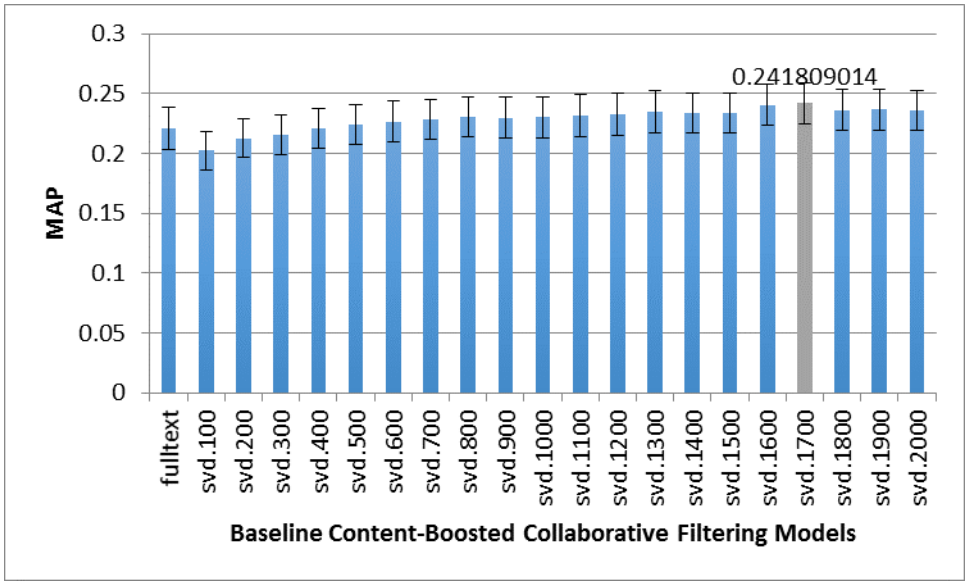


Figure 68: CBCF Baselines on 138 Homepage + Bibliography Fusion Users

The final comparison on the bottom right of Figure **67** showed that the individual SVD centroid model and the 15-NN.PO SVD model performed better than the K-Medoids SVD model but the result of centroid SVD model was slightly higher than the KNN.PO model. As a result, the individual SVD centroid model with 200 latent topics was chosen as the CBF baseline representative to test the hypothesis.

The CBCF baseline models with 138 users, who provided their homepage and publications, or had their homepages and publications identified by us, were assessed, as depicted in Figure **68**. Their MAP results increased gradually. The maximum MAP of CBCF baseline models was found in the CBCF SVD model, with 1700 latent topics.

The experimental CombMNZ and CombSUM fusion approaches, which fused the homepage-augmented models and the bibliography-augmented models, were assessed with 138 users who provided their homepages and publications, or had their homepages and publications identified by us, as shown in Figure **69**.

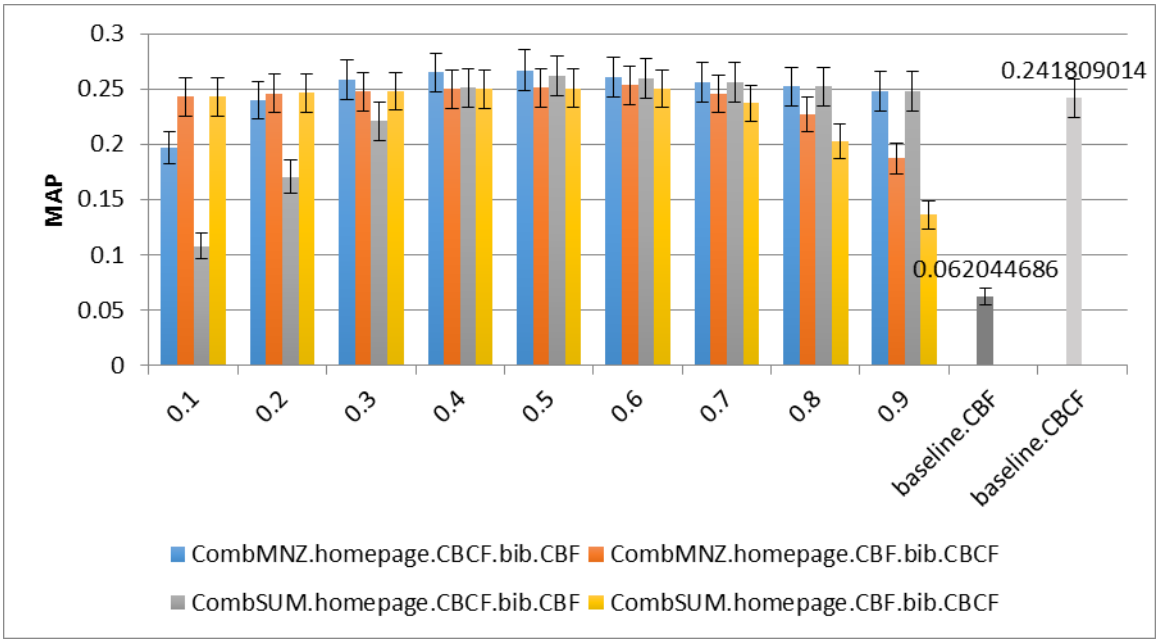


Figure 69: MAP results of Homepage + Bibliography Recommendation Fusion

The MAP results of the experimental homepage-augmented and bibliography-augmented recommendation fusion models varied from 0.1 to 0.27, depending on the type of fusion,

recommending method, and the weight of fusion. Models fusing with the CombMNZ method performed well in any stepping weights. The maximum MAP result of the CombMNZ fusion of homepage-augmented CBCF and bibliography-augmented CBF models at a weight 0.5 was 0.267. The peak MAP result of the CombMNZ models of the homepage-augmented CBF and the bibliography-augmented CBCF models at a weight of 0.6 was 0.253. On the other hand, the CombSUM models were quite sensitive to stepping weights. The poorest MAP results of the CombSUM models were 0.108 at the fusion of homepage-augmented CBCF and bibliography-augmented CBF models at a weight of 0.1, and 0.136 of the homepage-augmented CBF and the bibliography-augmented CBCF models at a weight of 0.9. However, the maximum MAP result of the CombSUM fusion of homepage-augmented CBCF and bibliography-augmented CBF models at a weight of 0.5 was 0.262. The peak MAP result of the CombSUM models of the homepage-augmented CBF and the bibliography-augmented CBCF models at a weight of 0.5 was 0.25.

In summary, as shown in Figure **69**, all 36 fusion models performed significantly better than the CBF baseline with a p-value < 0.01 significant level. However, only one model (the CombMNZ fusion of the homepage-augmented CBCF and the bibliography-augmented CBF models at a weight of 0.5) performed significantly better than the CBCF baseline with a p-value = 0.04842.

8.3.1.2 Homepage-External-bookmarks Fusion

The CBF baselines with the 30 users, who provided their homepage and external bookmark accounts or had these identified by us, were assessed as shown in Figure **70**. The maximum MAP result of the individual centroid models was 0.0747 on the unigram centroid model. This was the

peak MAP result compared to the other SVD counterparts, as shown in the top left diagram on Figure **70**. Other, later SVD CBF baseline models also used 400 latent topics. As shown in the top right of Figure **70**, all the clustering models performed quite at the same level.

The MAP result of K-Medoids SVD model was the highest among the clustering centroid models. As shown in the bottom left of Figure **70**, all the KNN.PO baseline models performed similarly with no significant differences among them. The maximum MAP result of the KNN.PO models was 0.073 at the 5-NN.PO full-text model.

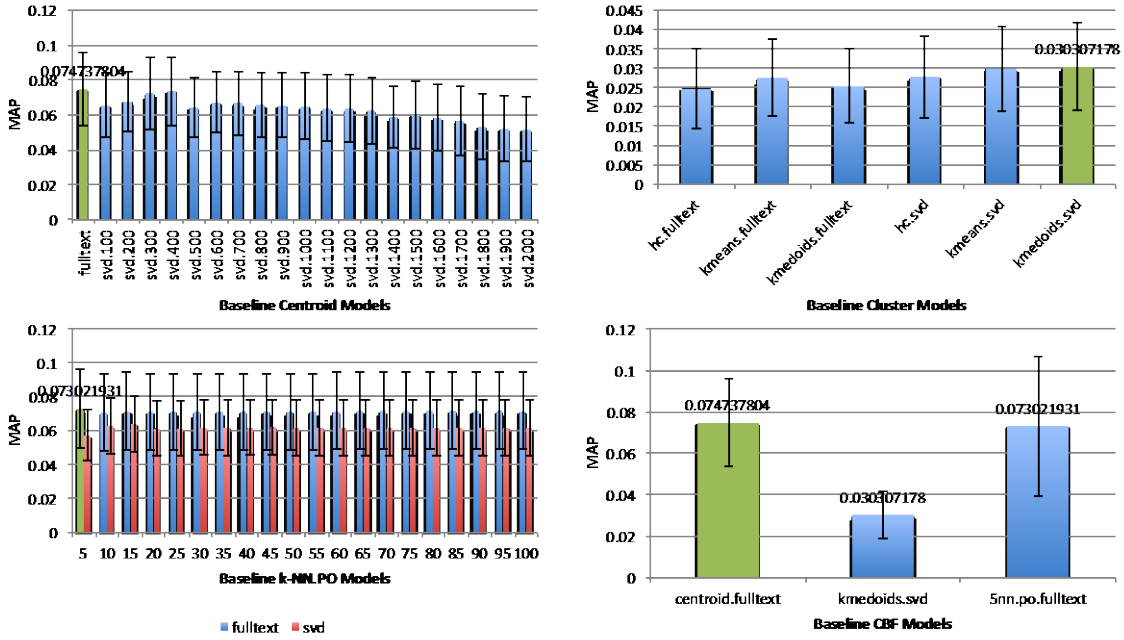


Figure 70: MAP Results of CBF Baselines on 30 Homepage + External Bookmark Fusion Users

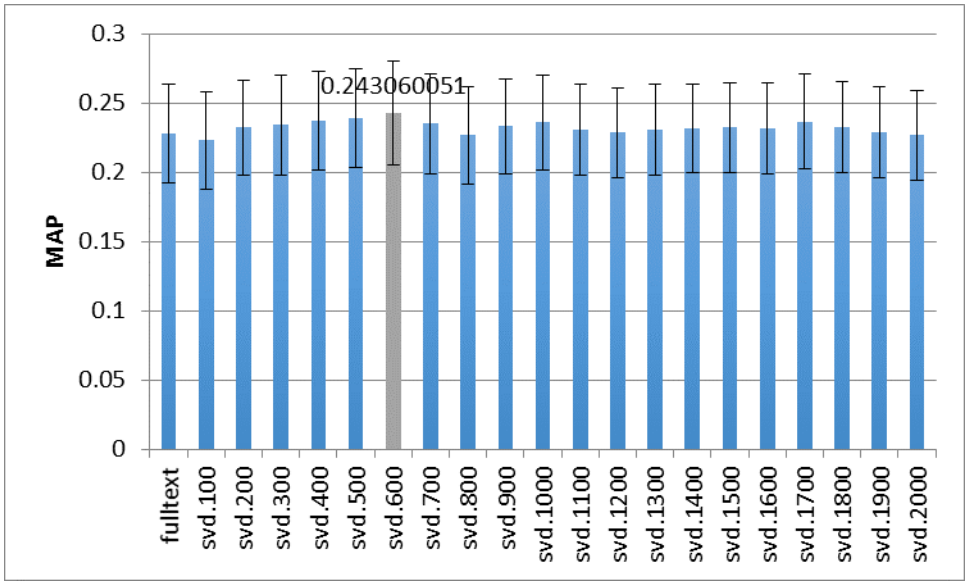


Figure 71: MAP Results of CBCF Baselines on 30 Homepage + External Bookmark Fusion Users

The final comparison on the bottom right of Figure **70** showed that the individual unigram centroid model and the 5-NN.PO unigram model performed better than the K-Medoids SVD model but the result of the unigram centroid model was slightly higher than the KNN.PO model. As a result, the individual unigram centroid model was chosen as the CBF baseline representative to test the hypothesis.

The CBCF baseline models with the same 30 users were assessed, as depicted in Figure **71**. Their MAP results stayed similarly and the maximum MAP of CBCF baseline models was found in the CBCF SVD model, with 600 latent topics.

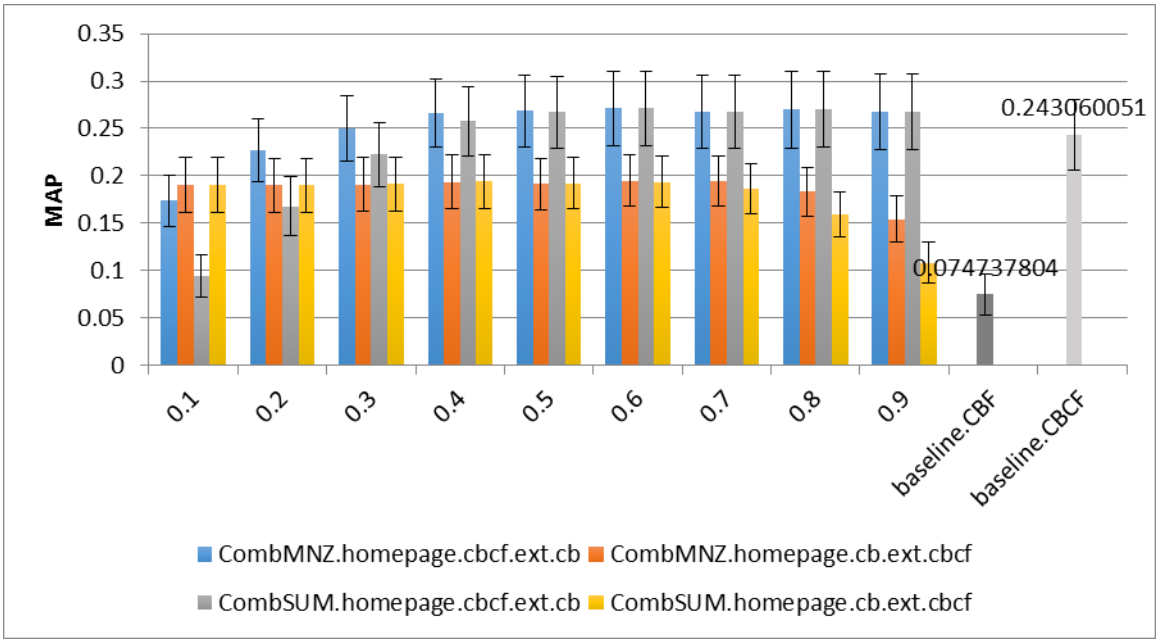


Figure 72: MAP Results of Homepage + External Bookmark Recommendation Fusion Models

The experimental CombMNZ and CombSUM fusion models with homepage and external scholarly bookmarked papers augmentations were assessed with 30 users, who provided their homepages and external scholarly bookmarked papers, or had these identified by us, as shown in Figure **72.**

The MAP results of the experimental homepage-augmented and external bookmark-augmented recommendation fusion models varied from 0.09 to 0.27, depending on the type of fusion, recommending method, and the weight of fusion. Models, which fused with both the CombMNZ and the CombSUM methods, were quite sensitive to stepping weights. The maximum MAP result of the CombMNZ fusion of homepage-augmented CBCF and external bookmark-augmented CBF models at a weight of 0.6 was 0.271. The peak MAP result of the CombMNZ models with homepage-augmented CBF and the external bookmark-augmented

CBCF models at a weight of 0.1 was 0.19. On the other hand, the poorest MAP results of CombMNZ models were 0.19 at the fusion of the homepage-augmented CBCF and the external bookmark-augmented CBF models at a weight of 0.1, and 0.15 of the homepage-augmented CBF and the external bookmark-augmented CBCF models at a weight of 0.9. On the CombSUM models, the maximum MAP result were 0.27 of the fusion model of the homepage-augmented CBCF and the external bookmark-augmented CBF models at a weight of 0.6, and 0.19 of the homepage-augmented CBF and the external bookmark-augmented CBCF models at a weight of 0.1. The minimum MAP results were 0.09 of the fusion model of the homepage-augmented CBCF and the external bookmark-augmented CBF models at a weight of 0.1, and 0.11 of the homepage-augmented CBF and the external bookmark-augmented CBCF models at a weight of 0.9.

In summary, from Figure **72**, 34 out of 36 fusion models (18 CombMNZ and 16 CombSUM models) performed significantly better than the CBF baseline with p-value < 0.01 significant levels and one CombSUM model did with a p-value < 0.05 significant level. However, none of them outperformed the CBCF baseline.

There were 13 models (seven CombMNZ and six CombSUM models) that performed slightly better than the CBCF baseline but the difference was not significant.

8.3.1.3 Bibliography-External-bookmark Fusion

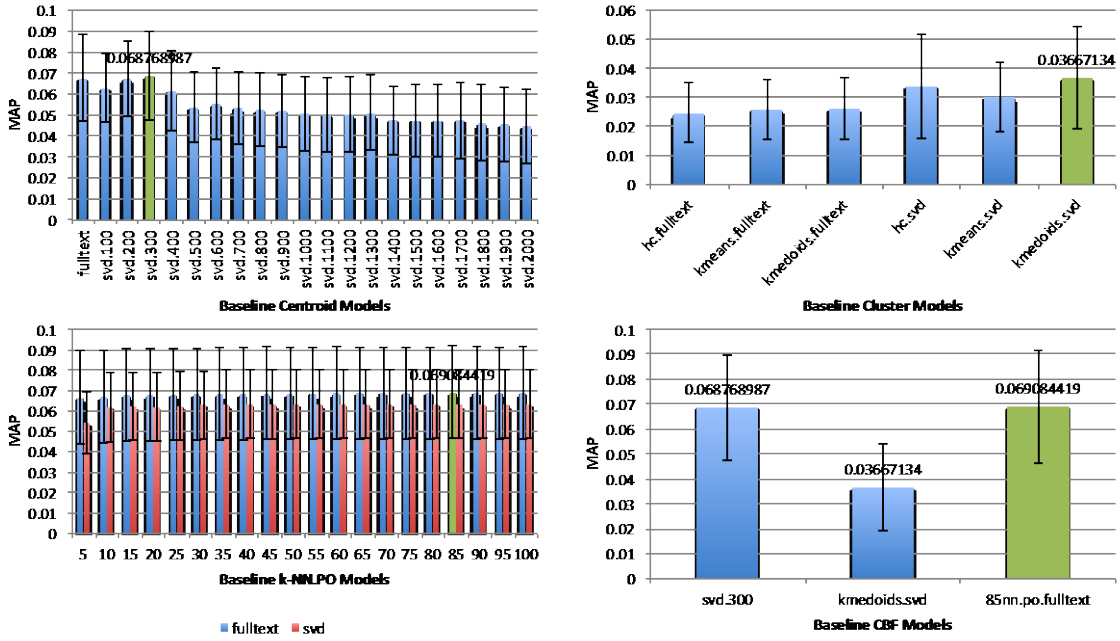


Figure 73: MAP Results of CBF Baselines on 27 Bibliography + External Bookmark Fusion Users

The CBF baselines with the 27 users, who provided their publications and external bookmark accounts or had these identified by us, were assessed, as shown in Figure **73**. The maximum MAP result of the individual centroid models was 0.0688 on the SVD centroid model, with 300 latent topics. This was the peak MAP result compared to the other SVD centroid and the unigram centroid models, as shown in the top left diagram on Figure **73**. Other, later SVD CBF baseline models used 300 latent topics. As shown in the top right of Figure **73**, all the three SVD clustering models performed slightly better than the unigram clustering centroid models. The MAP result of the K-Medoids SVD model was the highest among the clustering centroid models. As shown in the bottom left of Figure **73**, all the KNN.PO baseline models performed similarly with no

significant differences among them. The maximum MAP result of KNN.PO models was 0.0691 at the 85-NN.PO Full-text model.

The final comparison on the bottom right of Figure **73** showed that the individual SVD centroid model and the 85-NN.PO unigram model performed better than the K-Medoids SVD model, but the result of KNN.PO model was slightly higher than the SVD centroid model. As a result, the individual KNN.PO full-text model was chosen as the CBF baseline representative to test the hypothesis.

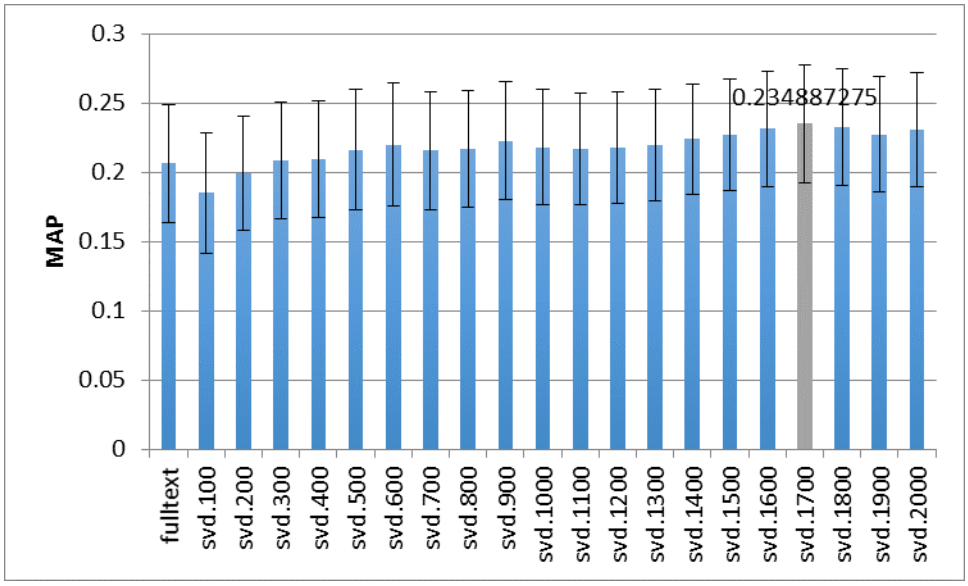


Figure 74: MAP results of CBCF Baselines on 27 Bibliography + External Bookmark Fusion Users

The CBCF baseline models with the same 27 users were assessed, as depicted in Figure **74**. Their MAP results were similar. The maximum MAP result of the CBCF baseline models was found in the CBCF SVD model, with 1700 latent topics.

The experimental CombMNZ and CombSUM fusion models with the bibliography and external scholarly bookmarked papers augmentations were assessed with 27 users, who provided their publications and external scholarly bookmarked papers, or had these identified by us, as shown in Figure **75.**

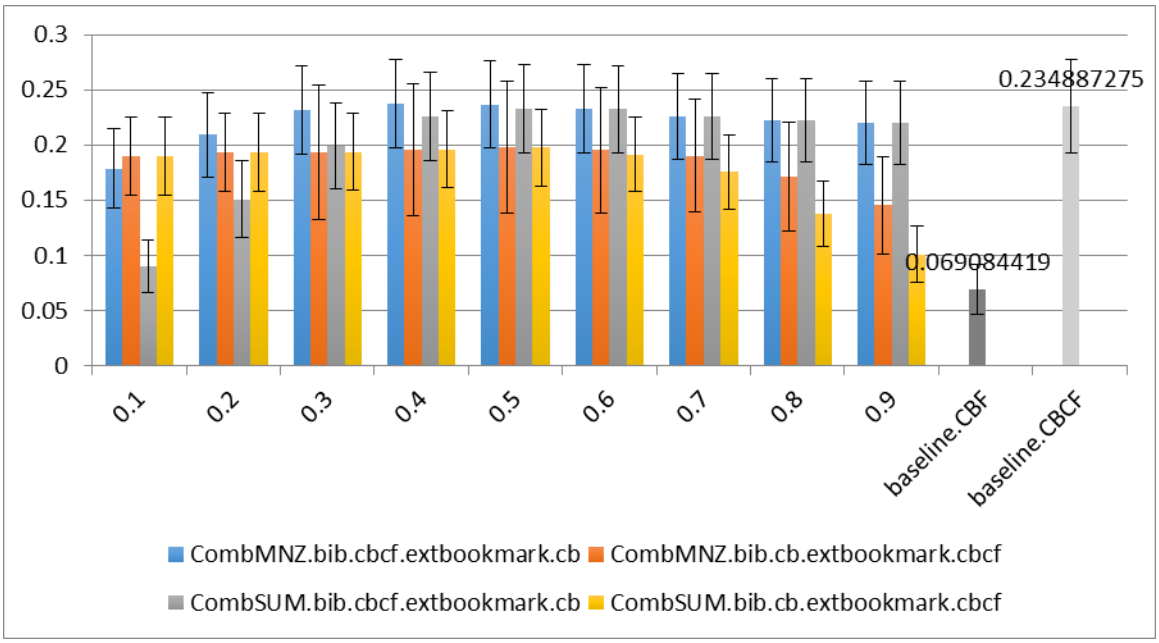


Figure 75: MAP Results of Bibliography + External Bookmark Recommendation Fusion Models

The MAP results of the experimental bibliography-augmented and external bookmark-augmented recommendation fusion models varied from 0.09 to 0.24, depending on the type of fusion, recommending method, and the weight of fusion. Models, which fused both the CombMNZ and the CombSUM methods, were quite sensitive to stepping weights. The maximum MAP result of the CombMNZ fusion of bibliography-augmented CBCF and external bookmark-augmented CBF models at a weight of 0.4 was 0.237. The peak MAP result of the

CombMNZ models of the bibliography-augmented CBF and the external bookmark-augmented CBCF models at a weight of 0.5 was 0.20. On the other hand, the poorest MAP results of CombMNZ models were 0.18 at the fusion of the bibliography-augmented CBCF and the external bookmark-augmented CBF models at a weight of 0.1, and 0.15 of the bibliography-augmented CBF and the external bookmark-augmented CBCF models at a weight of 0.9. On the CombSUM models, the maximum MAP result were 0.23 of the fusion model of the bibliography-augmented CBCF and the external bookmark-augmented CBF models at a weight of 0.5, and 0.2 of the bibliography-augmented CBF and the external bookmark-augmented CBCF models at a weight of 0.5. The minimum MAP results were 0.09 on the fusion model of the bibliography-augmented CBCF and the external bookmark-augmented CBF models at a weight of 0.1, and 0.1 of the bibliography-augmented CBF and the external bookmark-augmented CBCF models at a weight of 0.9.

In summary, as shown in Figure **75**, 34 out of 36 fusion models (18 CombMNZ and 16 CombSUM models) performed significantly better than the CBF baseline with p-value < 0.01 significance levels. However, none of them outperformed the CBCF baseline. There were two of them (two CombMNZ models) that performed slightly better than the CBCF baseline, but the difference was not significant.

8.3.2 Same-Source Fusion

8.3.2.1 Homepage

The homepage-augmented 20-NN.PO SVD CN3-term CBF model and the centroid SVD baseline CBF one with 100 latent topics were determined from Study 1. So did the homepage-

augmented SVD all-term CBCF model and the SVD baseline CBCF model, with 1600 latent topics. These four models were assessed with 317 CN3 users with homepages (either provided their homepage information to us directly or had their homepages identified by us).

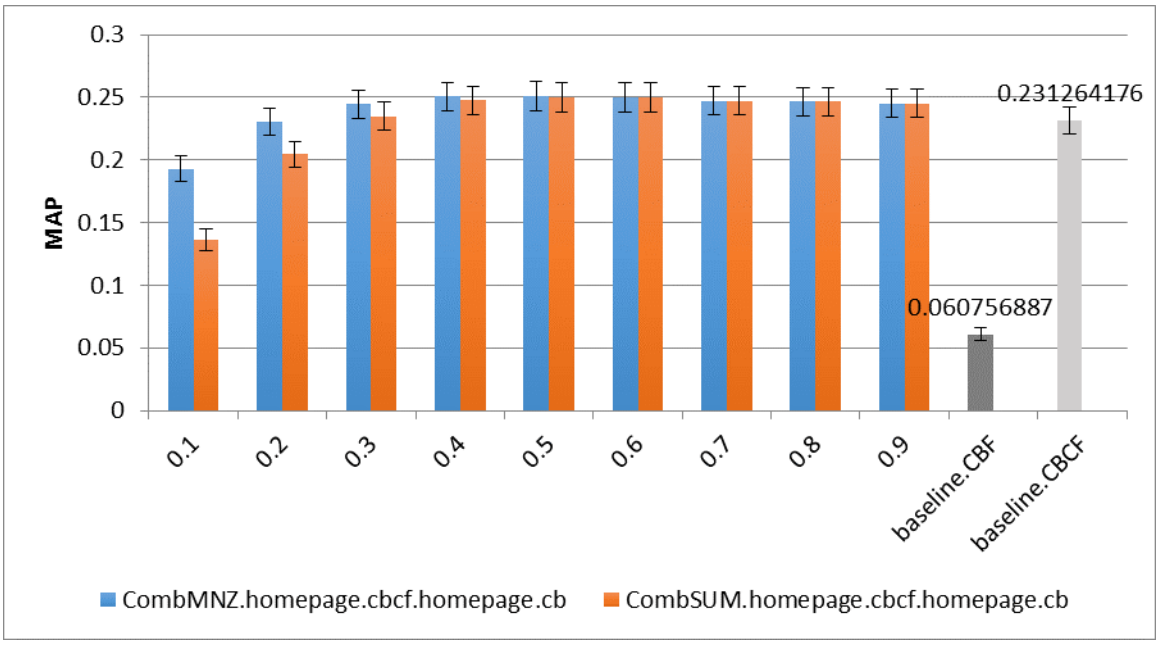


Figure 76: MAP Results of Homepage-Augmented Same-Source-Fusion Models

The MAP results of the experimental same-source-different-method homepage-augmented recommendation fusion models varied from 0.14 to 0.25, depending on the type of fusion, recommending method, and the weight of fusion. Models, which fused with the CombMNZ method, performed well on any stepping weights. The maximum MAP result of the CombMNZ fusion of the homepage-augmented models at a weight of 0.5 was 0.25. On the other hand, the CombSUM models were quite sensitive to stepping weights. The poorest MAP result

of the CombSUM models was 0.14 at a stepping weight of 0.1. However, the maximum MAP result of the CombSUM fusion model at a weight 0.5 was 0.25.

In summary, from Figure **76**, all 18 homepage-augmented fusion models performed significantly better than the CBF baseline with p-value < 0.01 significant levels. Also, eight models (four CombMNZ and four CombSUM models) performed significantly better than the CBCF baseline with the p-value < 0.05 significant levels.

8.3.2.2 User Publication (Bibliography)

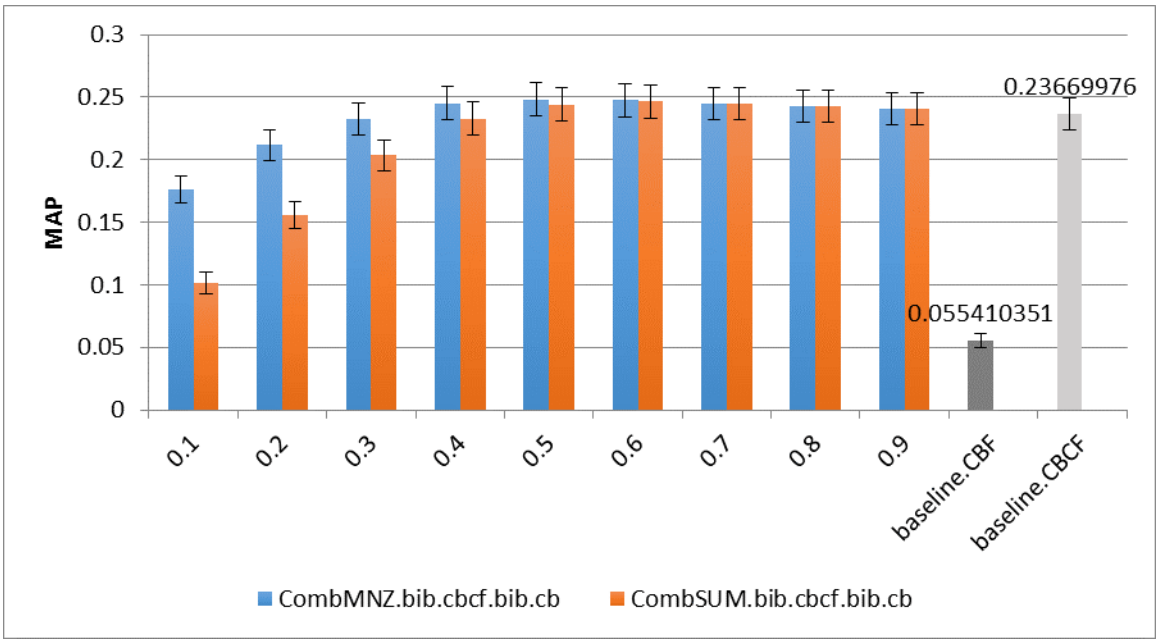


Figure 77: MAP Results of Bibliography-Augmented Same-Source-Fusion Models

The bibliography-augmented 5-NN.PO full-text CN3-term CBF model and the centroid SVD baseline CBF model, with 200 latent topics, were determined from Study 1. So did the

bibliography-augmented SVD CN3-term CBCF model, with 1900 latent topics, and the SVD baseline CBCF model, with 1700 latent topics. These four models were assessed with 249 CN3 users with bibliographies (either provided their publications directly to us or had their publications identified by us).

The MAP results of the experimental same-source-different-method bibliography-augmented recommendation fusion models varied from 0.1 to 0.25, depending on the type of fusion, recommending method, and the weight of fusion. Models, which fused with the CombMNZ method, performed well on any stepping weight. The maximum MAP result of the CombMNZ fusion of the bibliography-augmented models at a weight of 0.5 was 0.248. On the other hand, the CombSUM models were quite sensitive to stepping weights. The poorest MAP result of the CombSUM models was 0.1 at a stepping weight of 0.1. However, the maximum MAP result of the CombSUM fusion model at a weight of 0.6 was 0.246.

In summary, from Figure **77**, all 18 bibliography-augmented fusion models performed significantly better than the CBF baseline with p-value < 0.01 significance levels. However, none of them outperformed the CBCF baseline. While 11 models (six CombMNZ and five CombSUM models) performed slightly better than the CBCF baseline, the difference was not significant.

8.3.2.3 Bookmarked Scholarly Papers (External Bookmark)

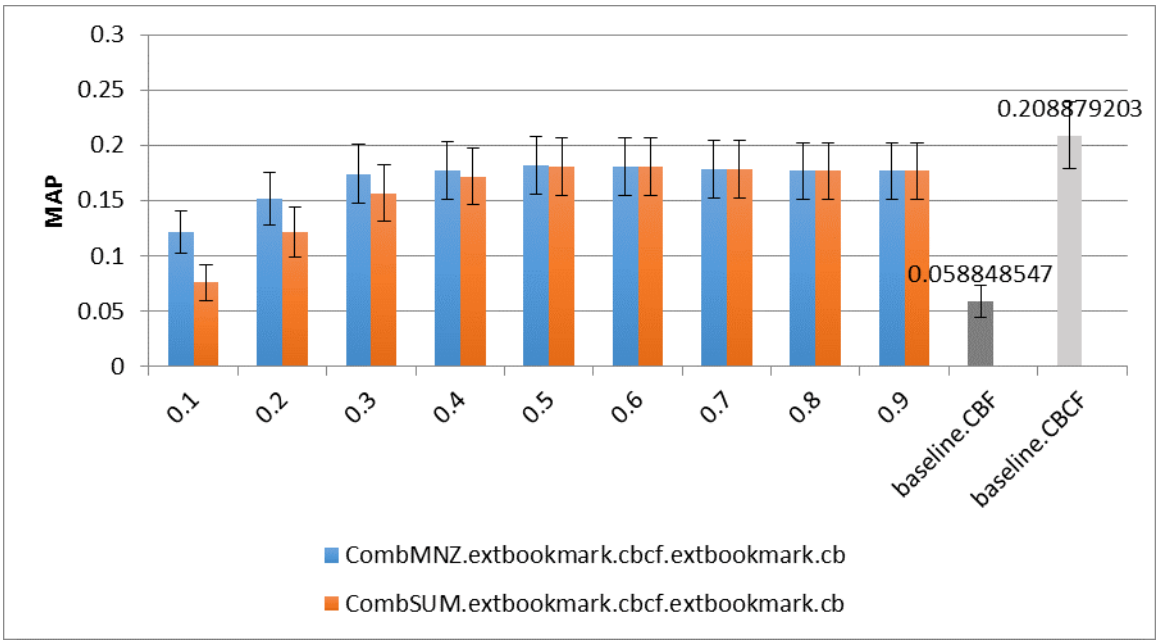


Figure 78: MAP Results of External Bookmark-Augmented Same-Source-Fusion Models

The external bookmark-augmented centroid full-text CN3-term CBF model and the centroid full-text baseline CBF model were determined from Study 1. So did the external bookmark-augmented SVD CN3-term CBCF model, with 2000 latent topics, and the SVD baseline CBCF model, with 1700 latent topics. These four models were assessed with 45 CN3 users with external bookmarks (either provided their external bookmark accounts directly to us or had their accounts identified by us).

The MAP results of the experimental same-source-different-method external bookmark-augmented recommendation fusion models varied from 0.08 to 0.18, depending on the type of fusion, recommending method, and the weight of fusion. Models, which fused with the

CombMNZ method, performed quite well on any stepping weight. The maximum MAP result of the CombMNZ fusion of the external bookmark-augmented models at a weight of 0.5 was 0.18. On the other hand, the CombSUM models were quite sensitive to stepping weights. The poorest MAP result of the CombSUM models was 0.08 at a stepping weight of 0.1. However, the maximum MAP result of the CombSUM fusion model at a weight 0.6 was 0.218.

In summary, as shown in Figure **78**, 17 external bookmark-augmented fusion models (nine CombMNZ and eight CombSUM models) performed significantly better than the CBF baseline with p-value < 0.01 significant levels. However, none out of them performed better than the CBCF baseline.

8.4 SUMMARY AND DISCUSSION

Out of 162, 158 (97.53%) different-approach CombMNZ or CombSUM recommendation fusion models, which consisted of 104 cross-source and 53 same-source models, outperformed the CBF baseline models significantly (157 models with p-value < 0.01 and one with p-value < 0.05 significant levels). Also, when comparing them to the CBCF baseline models, 9 (5.56%) different-approach CombMNZ or CombSUM recommendation fusions models (one cross-source and eight same-source models) significantly helped improve recommendations with p-value < 0.05 significant levels.

Among the cross-source fusion models, all models, which fused the homepage-augmented and bibliography-augmented recommendations on both the CombMNZ and the CombSUM methods, yielded significantly better results than the CBF baselines. Only one

CombMNZ model, which fused the homepage-augmented CBCF and the bibliography-augmented CBF models, with weight of 0.5 outperformed the CBCF baseline with a p-value = 0.04842 significant level. Out of 36, 34 fusion models (18 CombMNZ and 16 CombSUM models) of the homepage-augmented and the external bookmark-augmented models performed significantly better than the CBF baseline with p-value < 0.01 significant levels and one CombSUM model did with a p-value < 0.05 significant level. 34 fusion models (18 CombMNZ and 16 CombSUM ones) of the bibliography-augmented and the external bookmark-augmented model also significantly outperformed compared to the CBF baseline with p-value < 0.01 significance levels.

In the same-source fusion models, all the recommendation results from the fusion of the homepage-augmented and bibliography-augmented models, and almost all of the fusion of the external bookmark-augmented same-source-different-method models outperformed CBF baselines significantly with p-value < 0.01 significant levels. However, only eight same-source homepage-augmented fusion models outperformed the CBCF baseline model.

Overall, the CombMNZ fusion models were more reliable than the CombSUM models because the CombMNZ method depends on how many recommendation sources they get. In this case, in which there were only two recommendation sources, it confirmed that not only a high recommendation score, but also those they needed to come from both sources. Even though, only nine of the fusion models (one cross-source and eight same-source fusion models) outperformed the CBCF baseline models, the results suggested that there is room to improve by combing them with other advanced machine learning techniques such as multi-task learning.

9.0 EXTERNAL VALIDITY STUDY

In order to reconfirm the findings from the experiments 1 - 3, the 800 talks from the following six conferences, EC-TEL 2013, i-KNOW 2013, IUI 2014, UMAP 2014, Hypertext 2014, and EC-TEL 2014, were held and used to compare between the external-source-augmented approaches with the baselines from the previous findings.

9.1 DATASET DEMOGRAPHICS

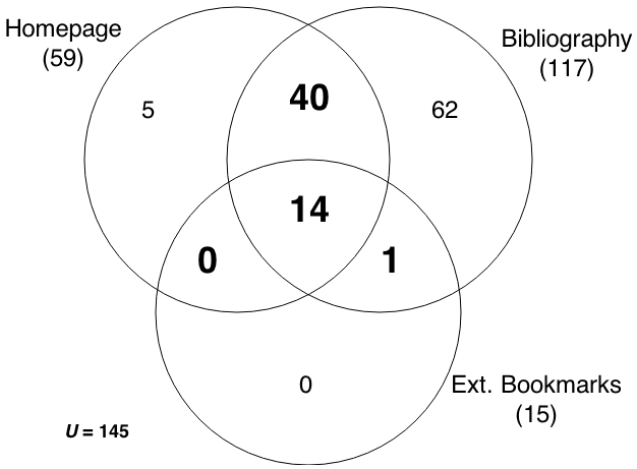


Figure 79: Users' Demography in CN3 in the External Validity for Reevaluation of the CN3 Studies

In the six holdout conferences, there were 145 users from the CN3 dataset bookmarking at least five talks. Of these, 122 users provided at least one external source, and also 55 users provided at least two external sources, as shown in Figure **79**.

These 122 qualified users for the reevaluation were those who bookmarked at least five talks in any of six holdout conferences, and who also provided at least one external source, consisting of: a personal webpage (59 users); user publication sources, including Google Citation (15 users) and Scopus (102 users); and/or external scholarly bookmark sources, including CiteULike (four users) and Mendeley (11 users) accounts. These sources were provided in the CN3 user personal information, or we identified their external sources proactively.

The distribution of the number of CN3 bookmarks per user is shown in Figure **80**. Out of 59 homepages that users provided, 42 (71.19%) had five pages or fewer. Out of the 117 users (74.36%) for whom we were able to identify and retrieve their publications, 87 had 20 publications or fewer. Of 15 users (53.33%) who provided external scholarly bookmark accounts, eight had 200 or fewer bookmarked papers in their accounts. However, seven users had more than 300 bookmarked articles (46.67%).

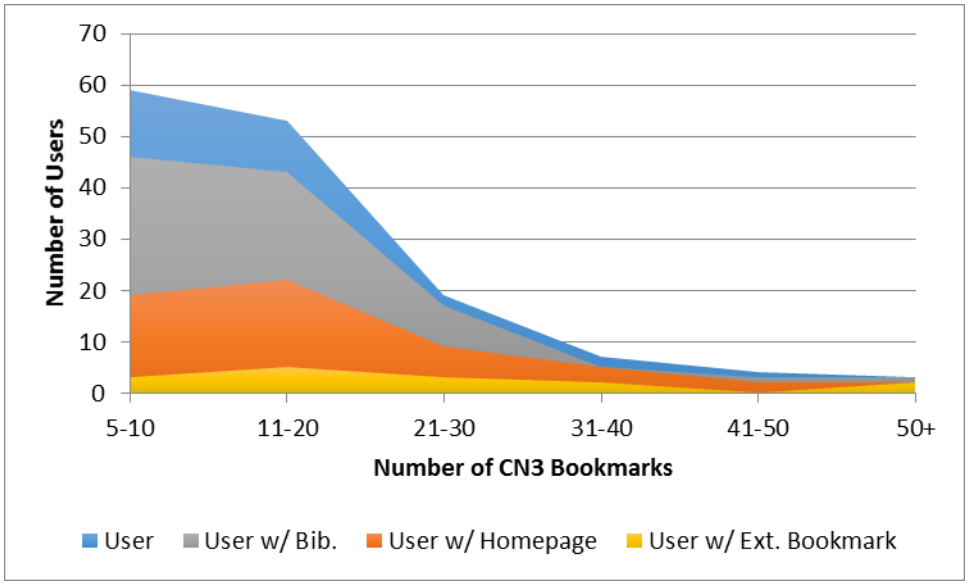


Figure 80: CN3 Bookmark Distributions in the Holdout CN3 Dataset

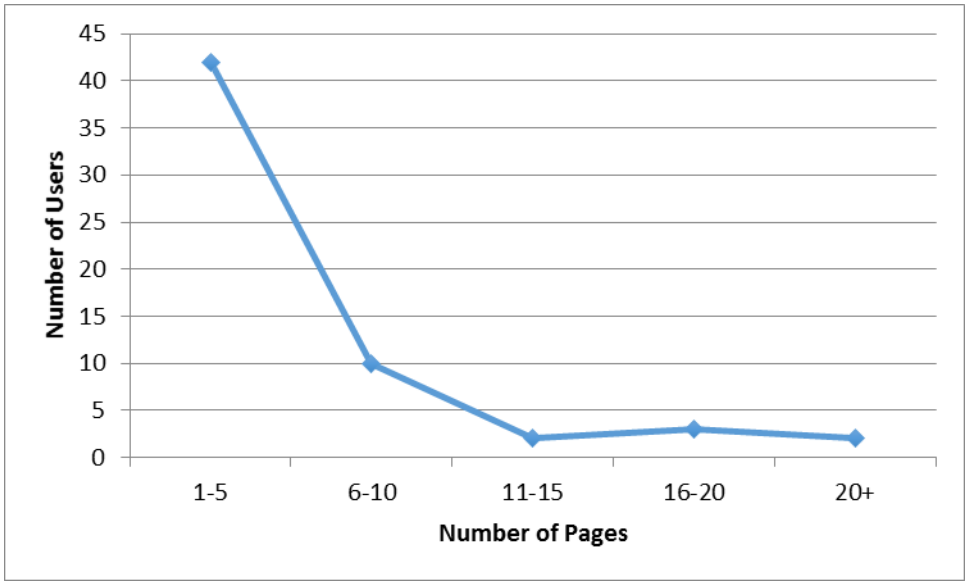


Figure 81: Homepage Distributions in the External Validity Analysis

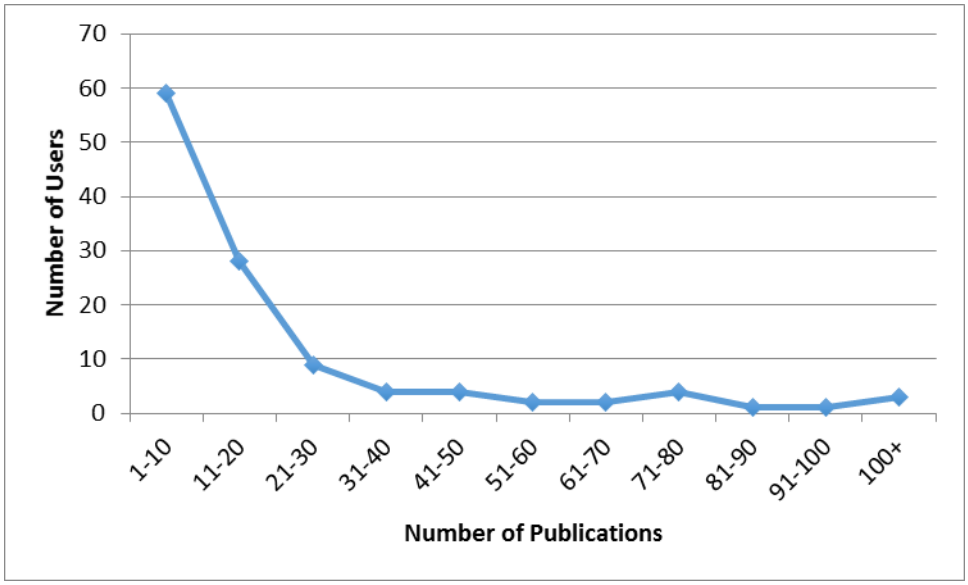


Figure 82: Bibliography Distributions in the External Validity Analysis

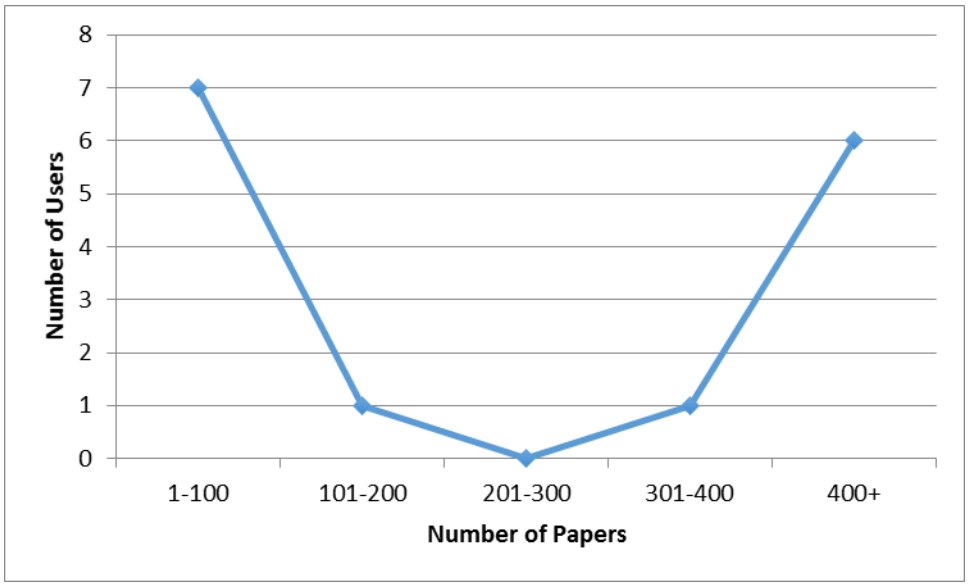


Figure 83: External Scholarly Bookmarked Papers Distribution in the External Validity Analysis

Five-fold cross-validations were conducted to evaluate the external validity of Study 1 and Study 3, and five-fold-ten-round cross-validation was conducted to evaluate the external validity

of Study 2. For the reassessment of Study 3, which concerned recommendation fusion, 55 out of 122 qualified users from the CN3 dataset had at least two external sources in the external validity dataset. These sources included: a personal webpage (54 users); user publication sources, including Google Citation (15 users) and Scopus (40 users), and external scholarly bookmark sources, including CiteULike (five users) and Mendeley (11 users), as shown in Figure **79** in bold face.

9.2 EXTERNAL VALIDITY OF STUDY 1: RECOMMENDATION IMPROVEMENT WITH EXTERNAL SOURCE AUGMENTATION

To assess the external validity of the study, six experimental (3 CBF + 3 CBCF) models and six baseline (3 CBF + 3 CBCF) models were selected from external-source-augmented models, in the same way as the ones from Study 1 were selected: recommendation improvement with external source augmentation. As Study 1 showed, cluster models and models running with all-term data performed poorly in content-based approaches. In this validity study, they were removed. Each model was run with the CN3-term matrix, except for the homepage-augmented CBCF models. The homepage-augmented CBCF models were run with the all-term TF-IDF matrix, because it was used in the selected model in Study 1.

Five-fold cross-validation was conducted to validate Study 3. For each user, the bookmarked talks from the CN3 holdout talks were split into five folds randomly. Each fold was used in turn for recommendation assessment. Talks in the first fold were combined with non-bookmarked talks as a test set. The remaining four folds were used to construct the experimental

recommendation approaches in each experiment. The evaluation result was the average of the results of five testing assessments.

9.2.1 Homepage

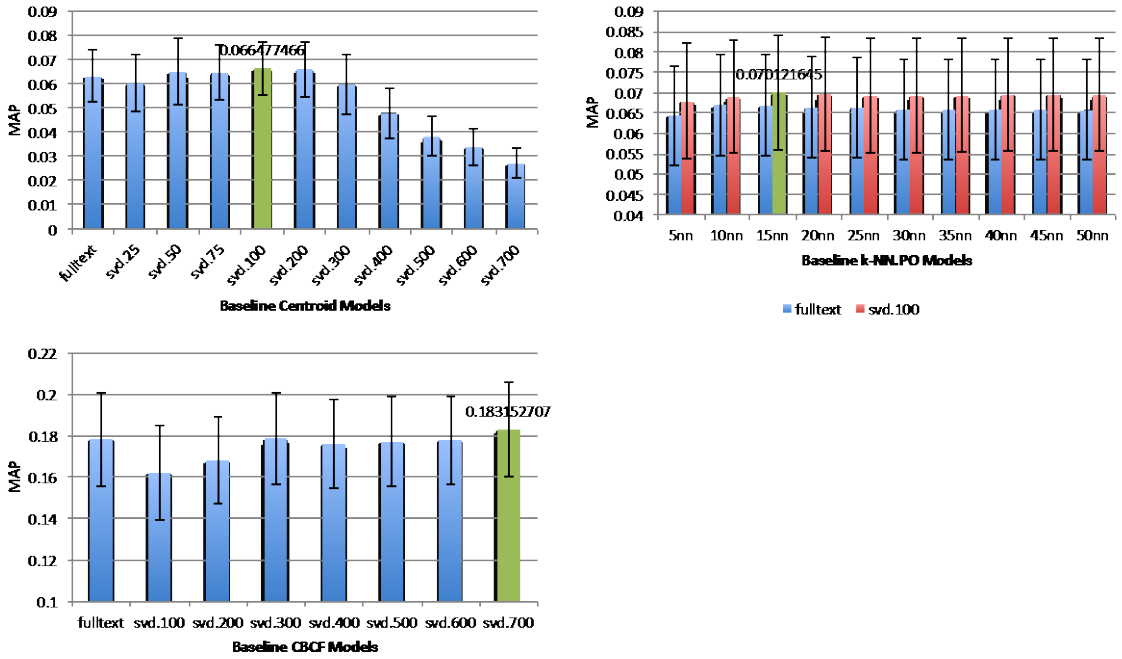


Figure 84: Baseline Models with Homepage Users on the External Validity on Study 1

The CBF baselines for the 59 users with a homepage (either provided to us directly or identified by us after the fact) were assessed for their external validity, as shown in Figure **84**. The maximum MAP result of individual centroid models was 0.066 on the SVD centroid model with 100 latent topics. The top left diagram in Figure **84** shows the peak MAP result in relation to the other SVD centroid and the full-text centroid models. the other SVD CBF baseline models also used 100 latent topics. As shown in the top right of Figure **84**, all the KNN.PO baseline models performed

similarly with no significant differences among them. The maximum MAP result of the KNN.PO model was 0.07 at the 15-NN.PO SVD model, with 100 latent topics.

The CBCF baseline models were assessed with 59 users, who provided their homepages or had their homepages identified by us, for their external validity as depicted in the bottom left of Figure **84**. Their MAP results performed similarly and the maximum MAP of CBCF baseline models was in the CBCF SVD model with 700 latent topics.

The 15-NN.PO SVD baseline model with 100 latent topics was selected as a content-based baseline model, among others as shown in the top left and top right charts of Figure **84**. The SVD content-boosted collaborative filtering baseline with 700 latent topics was also selected as a CBCF baseline representative shown in the bottom left of Figure **84**.

The homepage-augmented 35-NN.PO CN3-term SVD model with 200 latent topics was chosen as the candidate for the homepage-augmented CBF model, because its mean MAP result was the highest among the other content-based models shown in the top left and top right of Figure **85**. In the CBF comparison, both models were assessed with 59 users, who provided homepages and bookmarked CN3 talks in any of six holdout conferences. One-way Analysis of Variance (ANOVA) was applied to test the MAP results. From the top right of Figure **86**, the homepage-augmented CBF model performed slightly better than the CBF baseline did, but this difference was not statistically significant, because it failed to meet the standard of a p-value < 0.05 level for significance.

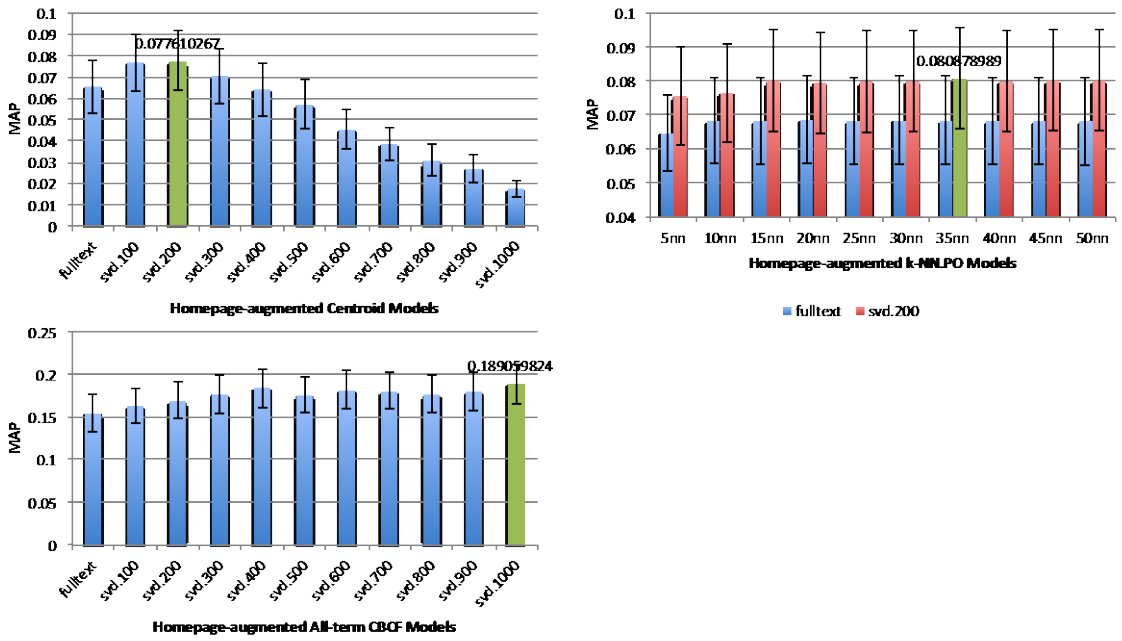


Figure 85: Homepage-Augmented Models of the External Validity of Study 1

In the CBCF comparison, the homepage-augmented all-term SVD content-boosted collaborative filtering model with 1000 latent topics, shown in the bottom left of Figure **85**, was assessed with the same 59 users. One-way ANOVA was applied to test the MAP results. From the bottom right of Figure **86**, the homepage-augmented CBCF model performed marginally better than the CBCF baseline, but once again failed to meet the standard of statistical significance, which was set at a p-value < 0.05.

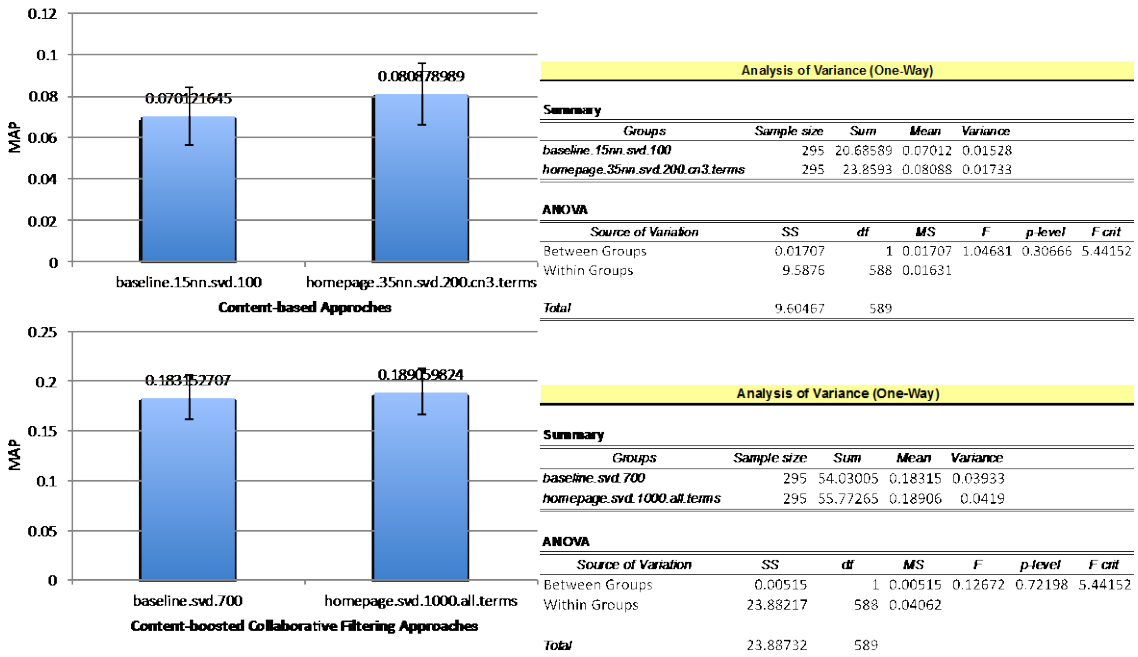


Figure 86: Hypothesis Testing Results for Homepage-Augmented Models on the External Validity of Study 1

9.2.2 Bibliography

The CBF baselines for the 117 users with publications (who either provided their publications to us or were identified by us) were assessed for their external validity, as shown in Figure 87. The maximum MAP result of the individual centroid models was 0.065 on the SVD centroid model with 200 latent topics. This was the peak MAP result in comparison with other SVD centroid and the full-text centroid models, as shown in the top left diagram on Figure 87. The baseline of 200 latent topics was also used later in the other SVD CBF baseline models. At the top right of Figure 87, all of the KNN.PO baseline models performed similarly with no significant differences among them. The maximum MAP result of the KNN.PO model was 0.07, using the 10-NN.PO SVD model with 200 latent topics.

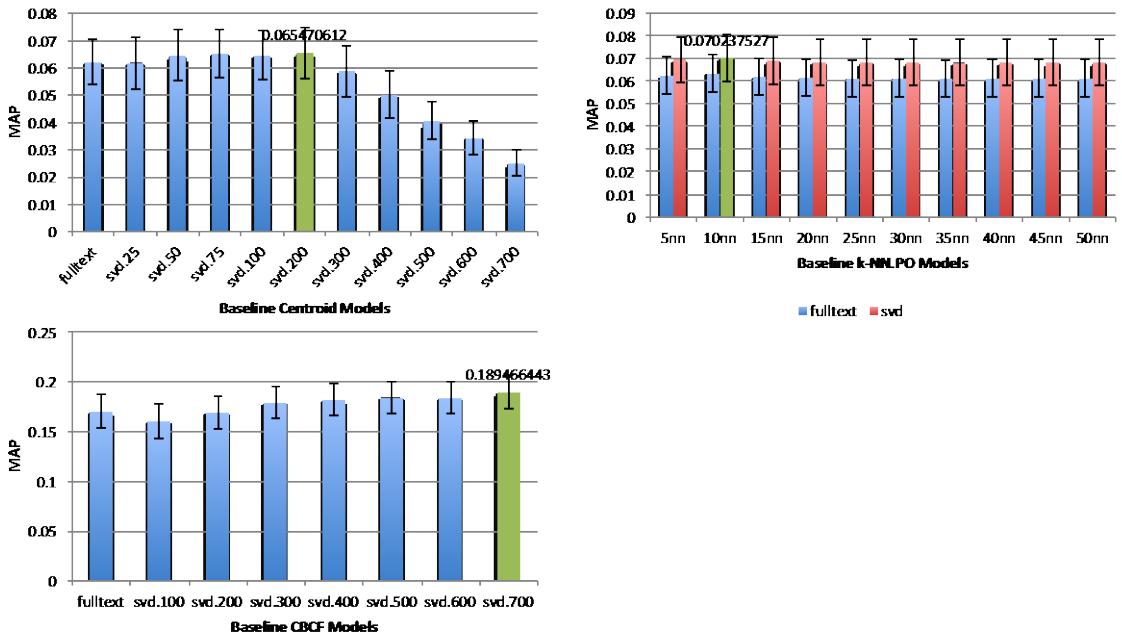


Figure 87: Baseline Models of Bibliography Users for the External Validity on Study 1

The CBCF baseline models were assessed for their external validity with 117 users, who either provided their publications or had their publications identified by us. The results are depicted in the bottom left of Figure 87. Their MAP results looked similarly and the maximum MAP result of the CBCF baseline models was seen in the CBCF SVD model with 700 latent topics.

The 10-NN.PO SVD baseline model with 200 latent topics was selected as a content-based baseline model, among others as shown in the top left and top right of Figure 87. The SVD content-boosted collaborative filtering baseline with 700 latent topics was also selected as a CBCF baseline representative, as shown in the bottom left of Figure 87.

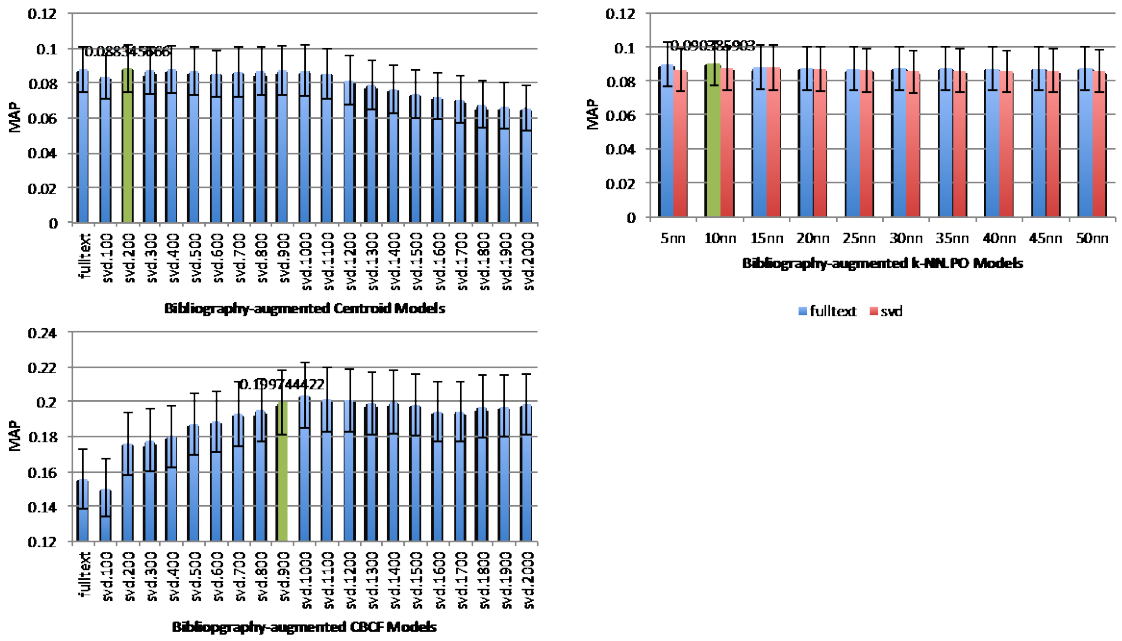
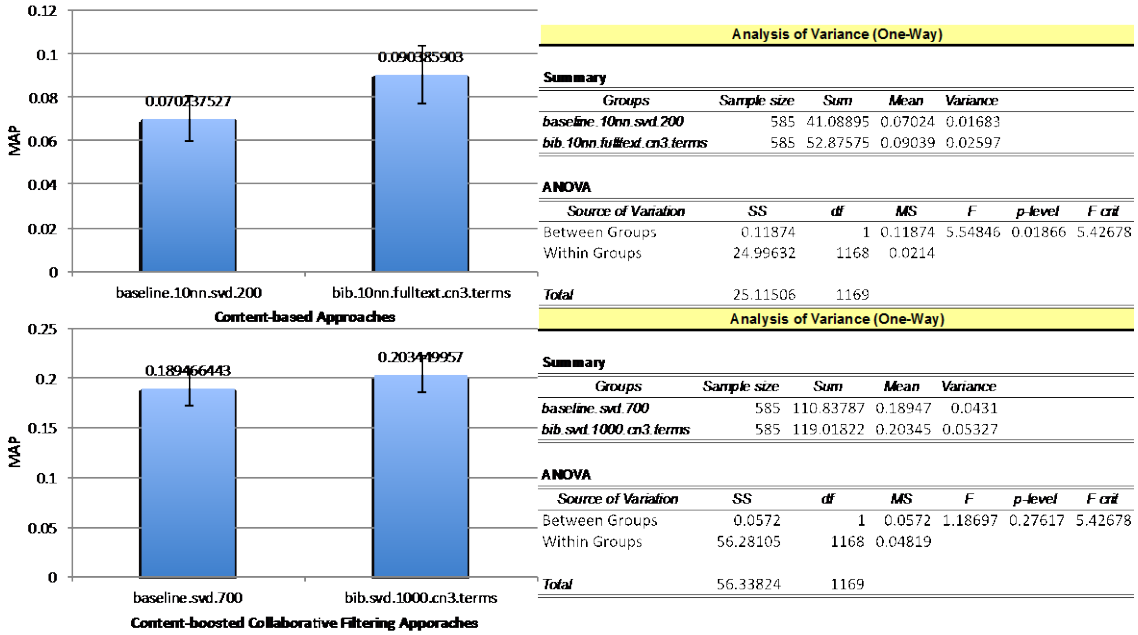


Figure 88: Bibliography-Augmented Models of the External Validity of Study 1

The bibliography-augmented 10-NN.PO CN3-term model, as shown in the top right of Figure **88**, was assessed with 117 users for whom we retrieved their publications, and who bookmarked CN3 talks in any of the six holdout conferences. One-way ANOVA was applied to test the MAP results. From the top right of Figure **89**, the bibliography-augmented CBF model performed better than the CBF baseline did, using a p-value < 0.05 standard for significance.

The bibliography-augmented SVD CN3-term content-boosted collaborative filtering model with 1000 latent topics, shown in the bottom left of Figure **88**, was assessed with the same 117 users against the SVD content-boosted collaborative filtering baseline with 700 latent topics, as shown in the bottom left of Figure 87. One-way ANOVA was applied to test the MAP results. From the bottom right of Figure **89**, the bibliography-augmented CBCF model performed slightly

better than the one at the CBCF baseline did, but the two models' performances were not statistically different, using a p-value < 0.05 level of significance.



Analysis of Variance (One-Way)

Summary

Groups	Sample size	Sum	Mean	Variance
baseline.10nn.svd.200	585	41.08895	0.07024	0.01683
bib.10nn.fulltext.cn3.terms	585	52.87575	0.09039	0.02597

ANOVA

Source of Variation	SS	df	MS	F	p-level	F crit
Between Groups	0.11874	1	0.11874	5.54846	0.01866	5.42678
Within Groups	24.99632	1168	0.0214			
Total	25.11506	1169				

Analysis of Variance (One-Way)

Summary

Groups	Sample size	Sum	Mean	Variance
baseline.svd.700	585	110.83787	0.18947	0.0431
bib.svd.1000.cn3.terms	585	119.01822	0.20345	0.05327

ANOVA

Source of Variation	SS	df	MS	F	p-level	F crit
Between Groups	0.0572	1	0.0572	1.18697	0.27617	5.42678
Within Groups	56.28105	1168	0.04819			
Total	56.33824	1169				

Figure 89: Hypothesis Testing Results for Bibliography-Augmented Models of the External Validity of Study 1

9.2.3 Bookmarked Scholarly Papers (External Bookmark)

The CBF baselines for the 15 users who either provided their external scholarly bookmark accounts, or had their accounts identified by us, were assessed for their external validity, as shown in Figure **90**. The maximum MAP result of individual centroid models was 0.079 on the SVD centroid model with 100 latent topics, which was the peak MAP result compared to other SVD centroid and the full-text centroid models, as shown in the top left diagram on Figure 87. Other SVD CBF baseline models also used 100 latent topics later. As shown in the top right of

Figure **90**, all the KNN.PO baseline models performed similarly with no significant difference among them. The maximum MAP result of KNN.PO model was 0.07, at the 15-NN.PO SVD model with 100 latent topics.

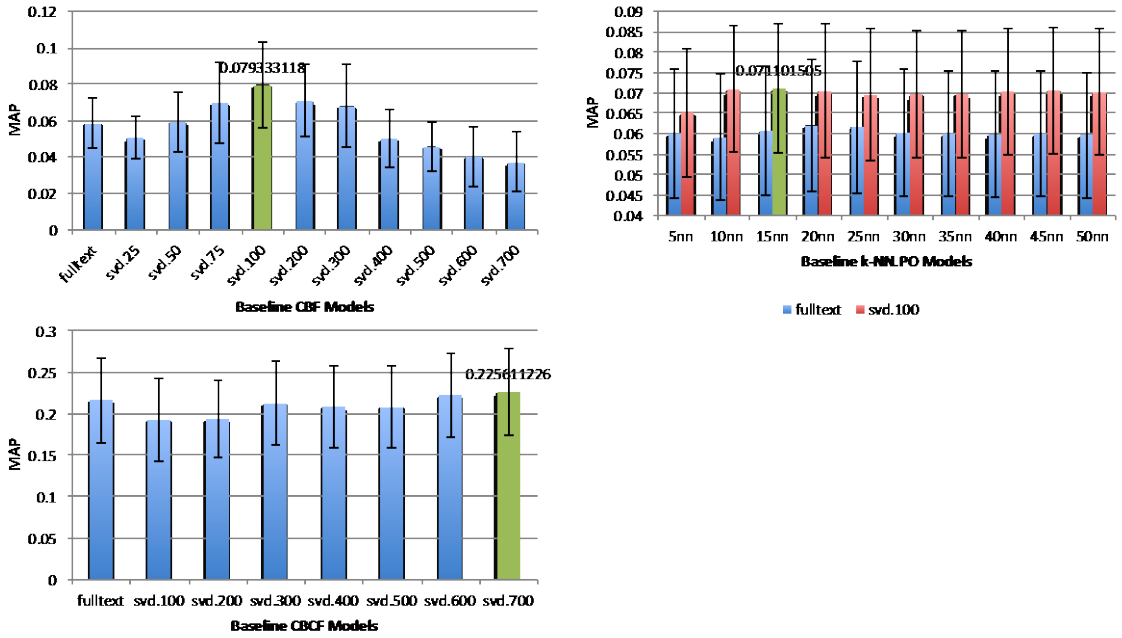


Figure 90: Baseline Models with External Bookmark Users for the External Validity of Study 1

The CBCF baseline models were assessed for their external validity with the 15 users who have their external scholarly bookmark accounts. This is depicted in the bottom left of Figure **90**. Their MAP results performed similarly, and the maximum MAP result of the CBCF baseline models was in the CBCF SVD model with 700 latent topics.

The centroid SVD baseline model with 100 latent topics was selected as a content-based baseline model, as shown in the top left and top right of Figure 90. The SVD CBCF baseline with

700 latent topics was also selected as a CBCF baseline representative, as shown in the bottom left of Figure 90.

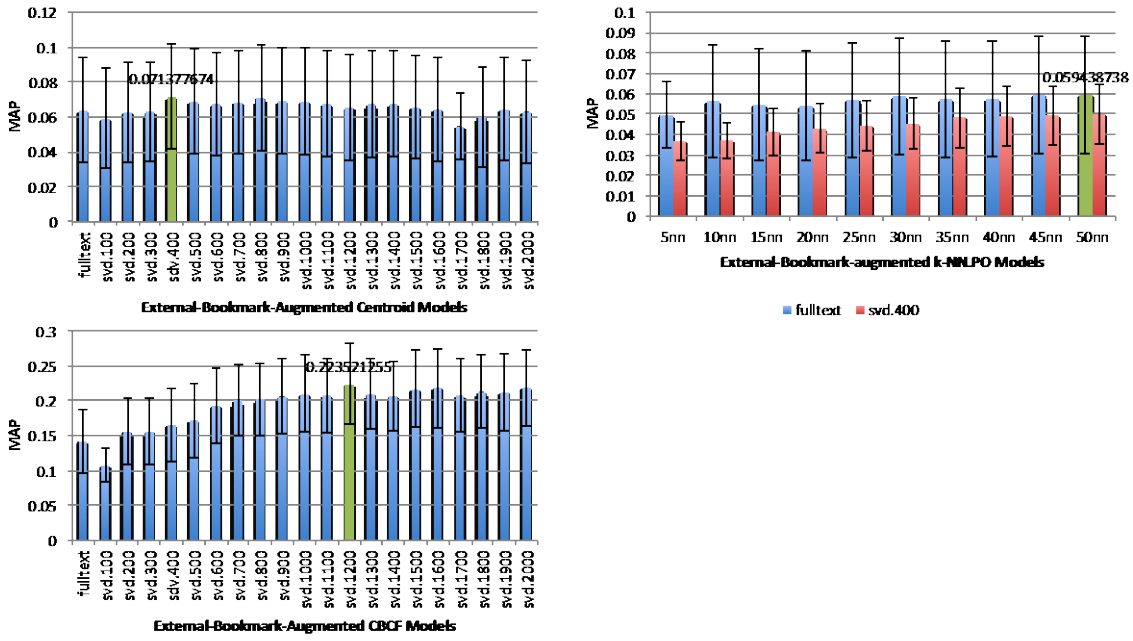
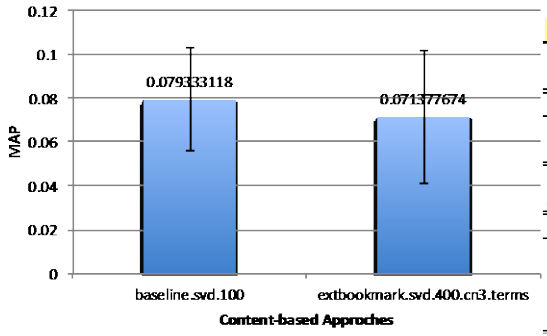


Figure 91: External Bookmark Models for the External Validity of Study 1

The external-bookmark-augmented SVD CN3-term centroid model with 400 latent topics, as shown in the top left of Figure 91, was assessed with 15 users, who provided their external scholarly bookmark accounts, and bookmarked talks in any of the six holdout conferences, against the SVD centroid baseline with 100 latent topics, as shown in the top left of Figure 90. One-way ANOVA was applied to test the MAP results. From the top right of Figure 92, the external-bookmark-augmented CBF models performed worse than the CBF baseline did, but there were no statistically significant difference, using a p-value < 0.05 standard for significance.

The external-bookmark-augmented CN3-term SVD CBCF model with 1200 latent topics, as shown in the bottom left of Figure 91, was assessed with the same 15 users against the SVD CBCF 100-latent-topic baseline, as shown in the bottom left of Figure 90. One-way ANOVA was applied to test the MAP results. From the bottom right of Figure 92, the external-bookmarked CBCF model performed marginally worse than the CBCF baseline did, but the difference was not statistically significant, using a p-value < 0.05 standard for significance.



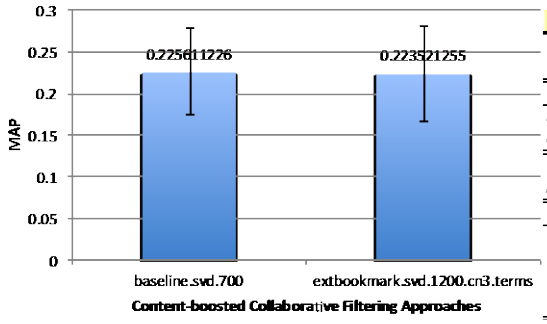
Analysis of Variance (One-Way)

Summary

Groups	*Sample size*	*Sum*	*Mean*	*Variance*
baseline.svd.100	75	5.94998	0.07933	0.01084
extbookmark.svd.400	75	5.35333	0.07138	0.01769

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*p-level*	*F crit*
Between Groups	0.00237	1	0.00237	0.16638	0.68394	5.531
Within Groups	2.11115	148	0.01426			
Total	2.11352	149				



Analysis of Variance (One-Way)

Summary

Groups	*Sample size*	*Sum*	*Mean*	*Variance*
baseline.svd.700	75	16.92084	0.22561	0.05335
extbookmark.svd.1200	75	16.76409	0.22352	0.06553

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*p-level*	*F crit*
Between Groups	0.00016	1	0.00016	0.00276	0.9582	5.531
Within Groups	8.79696	148	0.05944			
Total	8.79713	149				

Figure 92: Hypothesis Testing Results for External-Bookmark-Augmented Models of the External Validity of Study 1

9.3 EXTERNAL VALIDITY OF STUDY 2: COLD-START PROBLEM

To confirm the findings about the external-source-augmented models on the cold-start problem, the assessment of six experimental models against six baseline models was undertaken, as mentioned in the previous discussion of external validity.

To apply the same methodology to the cold-start effect study as described in section 7.2, the five-fold-ten-round cross-validation was conducted. In order to study the effect of cold-start problem, the size of bookmarked talks from the 807 CN3 holdout talks were divided into 21 different-sized bins, ranging from, no bookmark at all, to 20 bookmarks in the training set. The numbers of bins to which users were assigned was based on the number of bookmarks they had in their user profiles. For each user, the bookmarked talks were split into five folds randomly, as seen in Figure **61**a. Each fold in turn was used for recommendation assessment, along with the un-bookmarked talks. The remaining four folds were used to construct the experimental recommendation approaches. If the total bookmarked talks were more than 20, only 20 randomly selected talks were used for the assessment.

Secondly, as shown in Figure **61**b, the randomly picked talks in the training set were assigned to each cold-start-simulated window in accord to the cold-start situation. The random selection was repeated ten “round” times for every bin. The intermediate result for each bin was the average of these ten “round” iterations. The evaluation result was the average of the results of five testing assessments.

9.3.1 Homepage

In the CBF recommender type, the homepage-augmented 35-NN.PO SVD CN3-term CBF model with 200 latent topics was assessed in this cold-start problem against the 15-NN.PO SVD baseline CBF model with 100 latent topics. In the CBCF recommender type, the homepage-augmented SVD all-term CBCF model with 1000 latent topics was evaluated in this cold-start problem against the SVD baseline CBCF model with 700 latent topics.

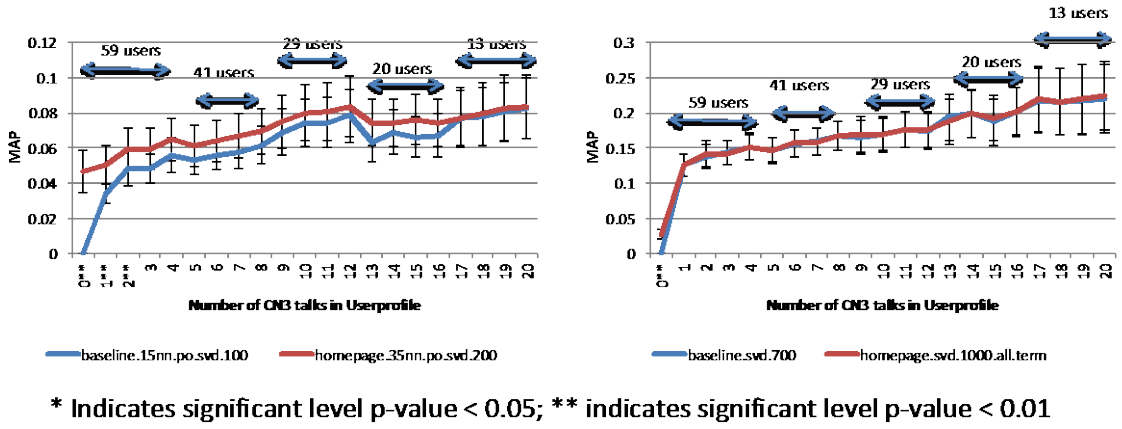


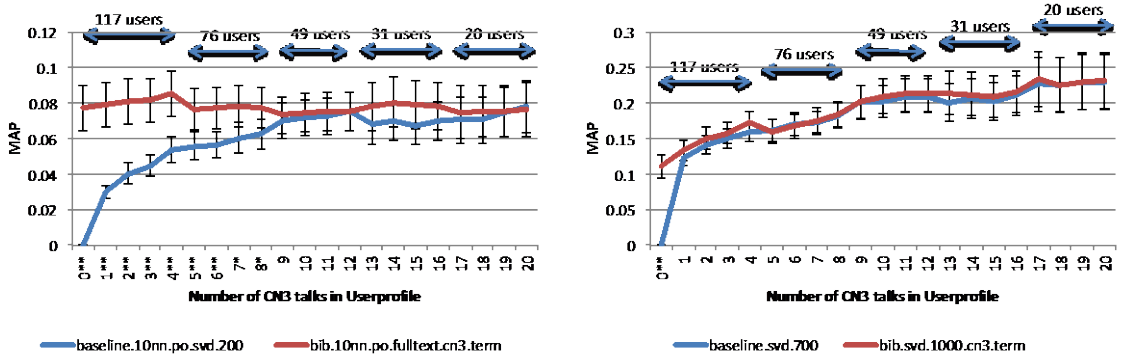
Figure 93: Homepage Models of the External Validity of the Cold-Start Study

These four models were assessed with 59 users, who provided their homepage information or had this information identified by us, and who also bookmarked talks in any of the six holdout conferences. One-way ANOVA was applied to test the MAP results. It was confirmed that the homepage augmentation recommendations on the CBF models improved significantly compared to the baselines in the initial cold-start situations, that users had no bookmark until there were two bookmarked talks. After that, the homepage-augmented CBF one performed slightly better

than the baselines but failed to meet the standard for statistical difference. The homepage-augmented CBCF model performed at the same level as the baseline model did in the most stages of the cold-start situations, except for the first stage, in which users had no bookmark.

9.3.2 Bibliography

In the CBF recommender type, the bibliography-augmented 10-NN.PO full-text CN3-term CBF model was assessed in the cold-start problem against the 10-NN.PO SVD baseline CBF model with 200 latent topics. In the CBCF recommender type, the bibliography-augmented SVD CN3-term CBCF model with 1000 latent topics was evaluated in the cold-start problem the SVD CBCF baseline model with 700 latent topics.



*** Indicates significant level p-value < 0.05; ** indicates significant level p-value < 0.01**

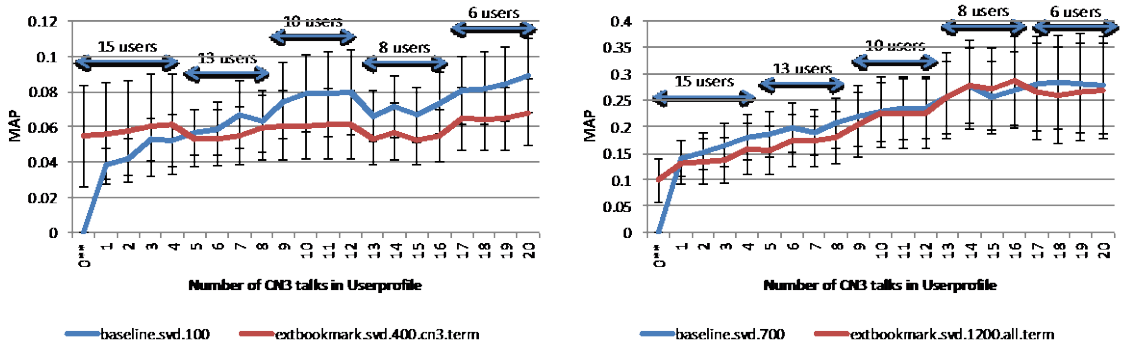
Figure 94: Bibliography Models of the External Validity of the Cold-Start Study

These four models were assessed with 117 users who provided their publications or had their publications identified by us, and who also bookmarked talks in any of the six holdout

conferences. One-way ANOVA was applied to test the MAP results. Like the homepage evaluation results, the bibliography-augmented recommendations from the CBF model outperformed the baseline models significantly in the no-bookmark cold-start situation, until there were eight bookmarked talks. After that, the bibliography-augmented CBF model performed slightly better than the baselines but the difference was not statistically different. The bibliography-augmented CBCF model performed at the same level as the baseline in the most stages of the cold-start situations, except in the first stage when users had no bookmarks.

9.3.3 Bookmarked Scholarly Papers (External Bookmark)

In the CBF recommender type, the external bookmark-augmented centroid SVD CN3-term CBF model with 400 latent topics was assessed in the cold-start problem against the centroid SVD baseline CBF model with 100 latent topics. In the CBCF recommender type, the external-bookmark-augmented SVD CN3-term CBCF model with 1200 latent topics was evaluated in the cold-start problem against and the SVD baseline CBCF model with 700 latent topics.



*** Indicates significant level p-value < 0.05; ** indicates significant level p-value < 0.01**

Figure 95: External Bookmark Models of the External Validity of Cold-Start Study

These four models were assessed with 15 users who provided their external bookmark accounts or had their accounts identified by us, and who also bookmarked talks in any of the six holdout conferences. One-way ANOVA was applied to test the MAP results. Recommendations with external bookmarked scholarly papers augmentation on both CBF and CBCF performed significantly better than the baselines at p-value < 0.05 level of significance in the initial stage of cold-start situations, when there were no bookmarked talks in the user profile. After that, the external bookmark-augmented CBF model performed slightly better in the early stages until there were five bookmarked talks, but then slightly underperformed compared to the CBF baseline. The external bookmark-augmented CBCF model performed at the same level as the CBF baseline did after the first stage.

9.4 EXTERNAL VALIDITY OF STUDY 3: RECOMMENDATION FUSION

To confirm the recommendation fusion findings of the external-source-augmented models, an assessment of six experimental models against six baseline models was conducted, as mentioned in the previous discussion of external validity.

Five-fold cross-validation was conducted for validating Study 3. For each user, the bookmarked talks from the CN3 holdout talks were randomly split into five folds. Each fold was used in turn for recommendation assessment. Talks in the first fold were combined with non-bookmarked talks as a test set. The remaining four folds were used for constructing the experimental recommendation approaches in each experiment. The evaluation result was the average of the results of the five testing assessments.

9.4.1 Different-Source-Different-Approach Fusion

9.4.1.1 Homepage-Bibliography Fusion

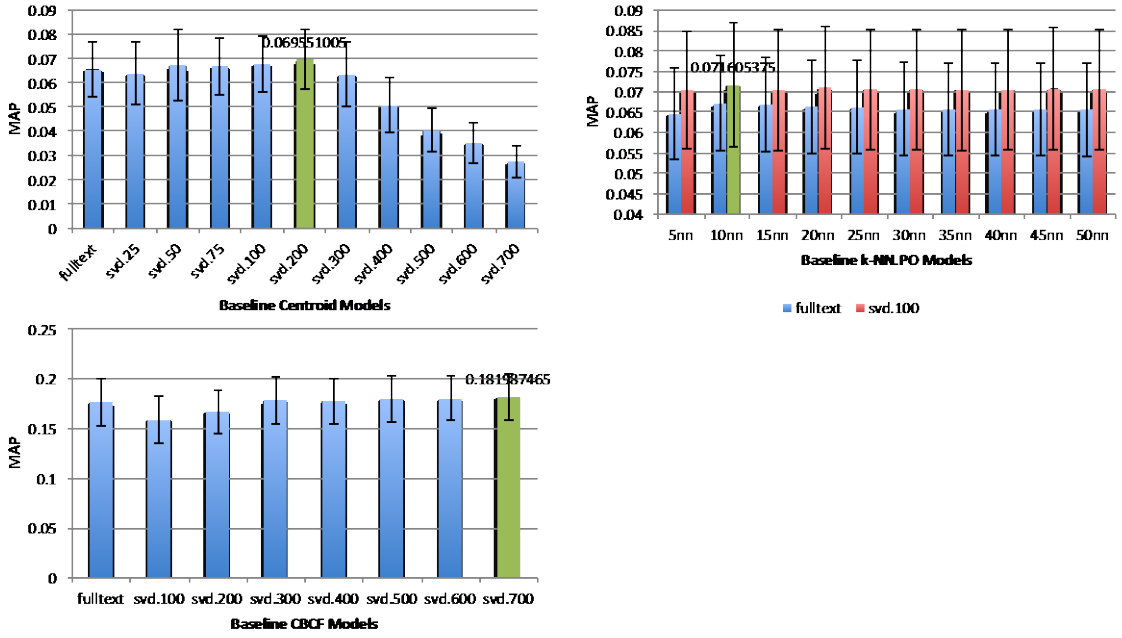


Figure 96: External Validity Baseline CBF on 54 Homepage + Bibliography Fusion Users

As shown in Figure **96**, the CBF baselines were assessed for their external validity with the 54 users, who provided their homepage and publications or had these identified by us. The maximum MAP result of individual centroid models was 0.0696 on the SVD centroid model with 200 latent topics, which was the peak MAP result compared to other SVD centroid and the full-text centroid models, as shown in the top left diagram in Figure **96**. Other, later SVD CBF baseline models also used 200 latent topics. As depicted in the top right of Figure **96**, all of the

KNN.PO baseline models performed similarly with no significant difference among them. The maximum MAP result of the KNN.PO model was 0.0716 for the 10-NN.PO SVD model.

The CBCF baseline models with 54 users, who provided their homepage and publications or had these identified by us, were assessed in the external validity as depicted in the bottom left of Figure **96**. Their MAP results looked similarly and the maximum MAP result of the CBCF baseline models was in the CBCF SVD model with 700 latent topics.

As determined from the external validity Study 1, the homepage-augmented 35-NN.PO SVD CN3-term CBF model with 200 latent topics and the homepage-augmented SVD all-term CBCF model with 1000 latent topics were the fusion models from the homepage source, and the bibliography-augmented 10-NN.PO full-text CN3-term CBF model and the bibliography-augmented SVD CN3-term CBCF model with 900 latent topics were the fusion models from the bibliography source. The 10-NN.PO SVD baseline model with 200 latent topics and the SVD CBCF baseline model with 700 latent topics were chosen as baselines for this evaluation.

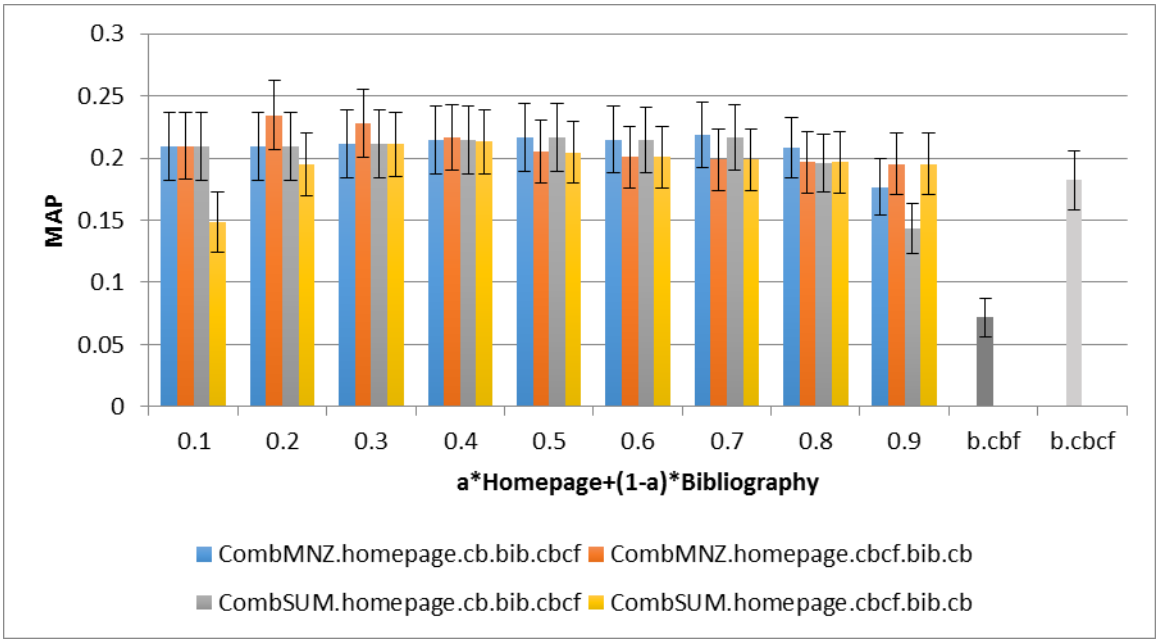


Figure 97: Homepage + Bibliography Fusion MAP of the External Validity

The experimental CombMNZ and CombSUM fusion approaches, which fused the homepage-augmented models and bibliography-augmented ones, were assessed with the 54 users who provided their homepages and publications, or had these identified by us, and who bookmarked talks in any of the six holdout conferences, as shown in Figure **97.**

One-way ANOVA with Tuky's HSD Post Hoc test was applied to test the MAP results. The MAP results of experimental homepage-augmented and bibliography-augmented recommendation fusion models varied from 0.14 to 0.23, depending on the type of fusion, recommending method, and the weight of fusion. Models fusing with the CombMNZ method performed well in any stepping weight. The range of MAP results of the CombMNZ fusions of the homepage-augmented and the bibliography-augmented models ranged between 0.18 and 0.23, and the peak result of the CombMNZ of the homepage-augmented CBCF and the

bibliography-augmented CBF models at a weight of 0.2 was 0.234. However, the CombSUM models were quite sensitive to stepping weights. The poorest MAP results of the CombSUM models were 0.15 at the fusion of homepage-augmented CBCF and bibliography-augmented CBF models at a weight of 0.1, and 0.14 of the homepage-augmented CBF and the bibliography-augmented CBCF models at a weight of 0.9, respectively. However, the maximum MAP result of the CombSUM fusion of the homepage-augmented CBCF and the bibliography-augmented CBF models at a weight of 0.4 was 0.213 and the peak result of the CombSUM of the homepage-augmented CBF and the bibliography-augmented CBCF models at a weight 0.7 was 0.217.

In summary, all 36 homepage-bibliography-fusion models (18 CombMNZ and 18 CombSUM models) significantly outperformed the CBF baseline at a p-value < 0.05 level of significance. However, only two (the CombMNZ of the homepage-augmented CBCF and the bibliography-augmented CBF models at weights of 0.2 and 0.3) significantly outperformed the CBCF baseline at the p-value < 0.05 level of significance.

9.4.1.2 Homepage-External-Bookmark Fusion

As shown in Figure **98**, the CBF baselines were assessed for their external validity with 14 users, who provided their homepage and external scholarly bookmark accounts or whose homepage and scholarly bookmark accounts were identified by us. The maximum MAP result of the individual centroid models was 0.069 on the SVD centroid model with 100 latent topics. This was the peak MAP result compared to the other SVD centroid and the full-text centroid models, which are shown in the top left diagram on Figure **98**. Other, later SVD CBF baseline models also used 100 latent topics. As shown in the top right of Figure **98**, all the KNN.PO baseline models

performed similarly with no significant differences in their performance. The maximum MAP result of the KNN.PO model was 0.0639 at the 20-NN.PO full-text CBF model.

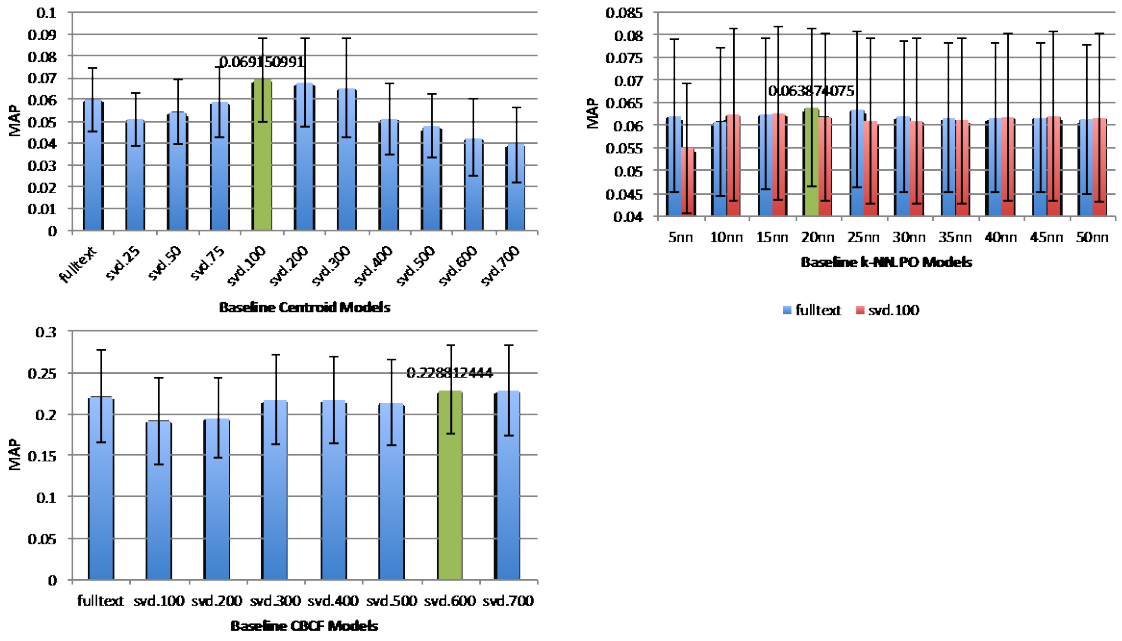


Figure 98: External Validity Baseline CBF for 14 Homepage + External Bookmark Fusion Users

The CBCF baseline models with 14 users, who provided their homepage and scholarly bookmark accounts or whose homepage and scholarly bookmark accounts were identified by us, were assessed for their external validity as depicted in the bottom left of Figure **98**. Their MAP results performed steadily and the maximum MAP result of the CBCF baseline models was found in the CBCF SVD model with 600 latent topics.

As determined from the external validity of Study 1, the homepage-augmented 35-NN.PO SVD CN3-term CBF model with 200 latent topics and the homepage-augmented SVD all-term CBCF model with 1000 latent topics were the fusion models from the homepage source.

The external-bookmark-augmented centroid SVD CN3-term CBF model with 400 latent topics and the external-bookmark-augmented SVD CN3-term CBCF model with 1,200 latent topics were the fusion models from the external bookmark sources. The centroid SVD baseline model with 100 latent topics and the SVD CBCF baseline model with 600 latent topics were chosen as baselines for this evaluation.

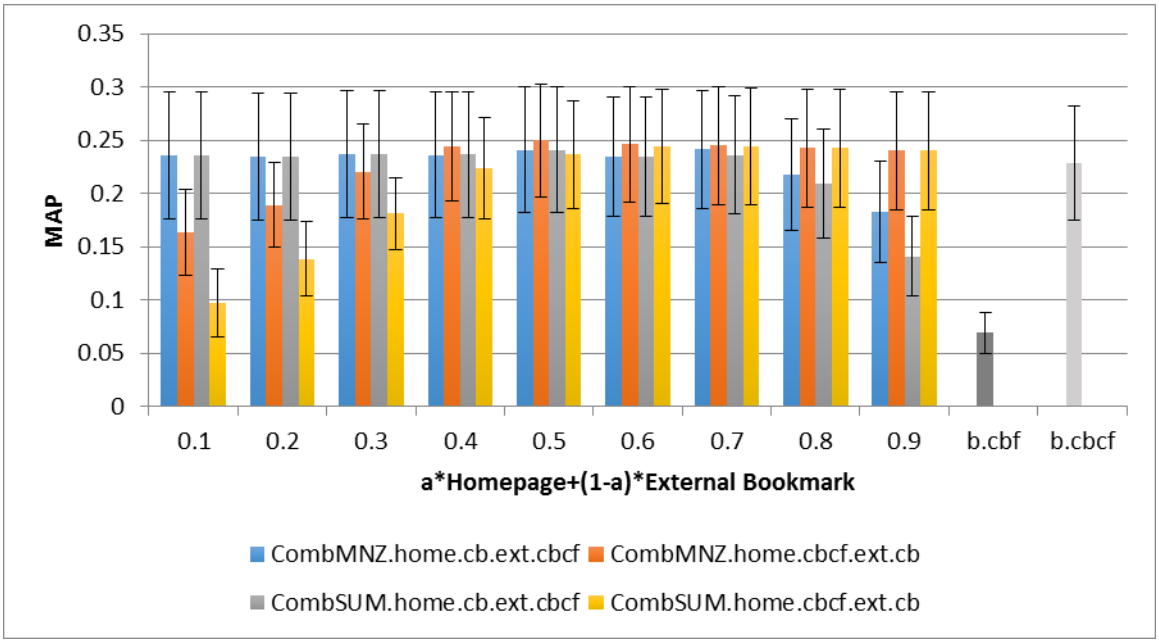


Figure 99: Homepage + External Bookmark Fusion MAP of the External Validity

The experimental CombMNZ and CombSUM fusion approaches, which fused between homepage-augmented models and external-bookmark-augmented ones, were assessed with 14 users who provided their homepages and external scholarly bookmark accounts, and bookmarked talks in any of the six holdout conferences, as shown in Figure 99.

One-way ANOVA with Tuky's HSD Post Hoc test was applied to test the MAP results. The MAP results of experimental homepage-augmented and external-bookmark-augmented recommendation fusion models varied from 0.1 to 0.25, depending on the type of fusion, recommending method, and the weight of fusion. Models fusing with the CombMNZ method were quite sensitive to stepping weights. The poorest MAP results of the CombMNZ models were 0.18 at the fusion of the homepage-augmented CBCF and the external-bookmark-augmented CBF models at a weight of 0.9, and 0.16 of the homepage-augmented CBF and the external-bookmark-augmented CBCF models at a weight of 0.1, respectively. However, the maximum MAP result of the CombMNZ fusion of the homepage-augmented CBCF and the external-bookmark-augmented CBF models at a weight of 0.5 was 0.25. The peak MAP result of the CombMNZ fusion of the homepage-augmented CBF and the external-bookmark-augmented CBCF models at a weight of 0.7 was 0.24. The CombSUM models were quite sensitive to stepping weights. The poorest MAP results of the CombSUM models were 0.14 at the fusion of the homepage-augmented CBCF and the external-bookmark-augmented CBF models at a weight of 0.1, and 0.1 of homepage-augmented CBF and external-bookmark-augmented CBCF models at a weight of 0.9. However, the maximum MAP result of the CombSUM fusion between the homepage-augmented CBCF and the external-bookmark-augmented CBF models at a weight of 0.7 was 0.24. The peak MAP result of the CombSUM fusion of the homepage-augmented CBF and the external-bookmark-augmented CBCF models at a weight of 0.5 was 0.24.

In summary, from Figure 99, 34 fusion models (18 CombMNZ and 14 CombSUM models) significantly outperformed the CBF baseline at the p-value < 0.05 level of significance. However, none of them outperformed the CBCF baseline.

9.4.1.3 Bibliography-External-Bookmark Fusion

The CBF baselines were assessed for their external validity with 15 users, who provided their publications and external scholarly bookmark accounts, or had these accounts and publications identified by us. The results of this assessment are shown in Figure 100. The maximum MAP result of the individual centroid models was 0.0793 on the SVD centroid model with 100 latent topics. This was the peak MAP result compared to the other SVD centroid and the full-text centroid models, as shown in the top left diagram on Figure 100. Other SVD CBF baseline models also used 100 latent topics. On the top right of Figure 100, all the KNN.PO baseline models performed similarly, with no significant differences among them. The maximum MAP result of the KNN.PO model was 0.0711 at the 15-NN.PO SVD model.

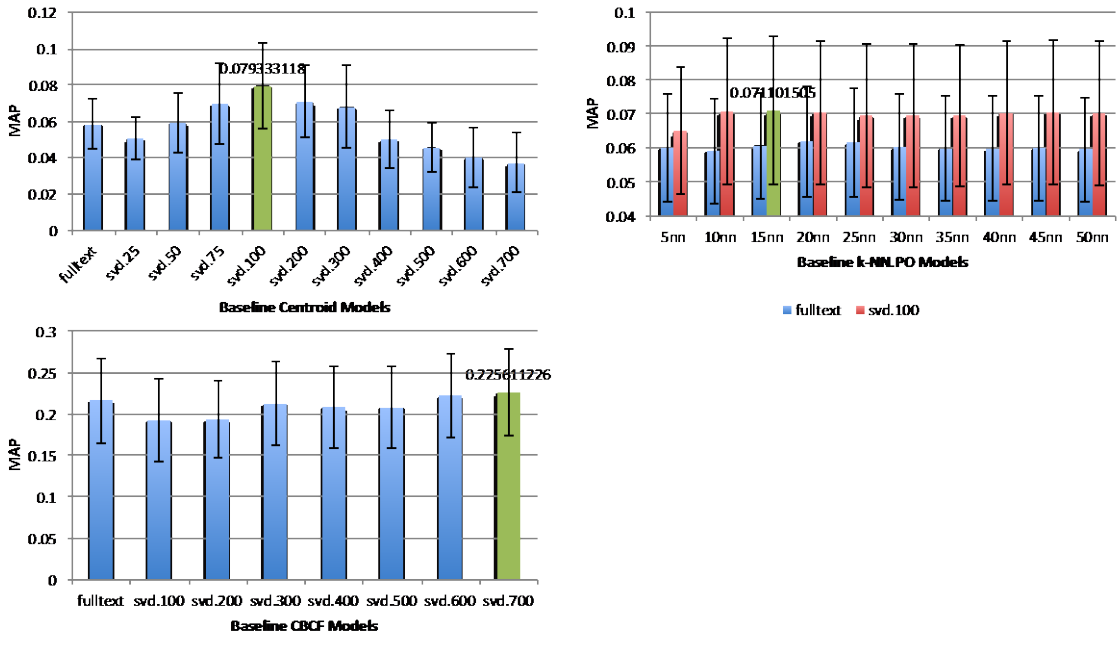


Figure 100: External Validity Baseline CBF for 15 Bibliography + External Bookmark Fusion Users

The CBCF baseline models were assessed for their external validity with 15 users, who provided their homepage and publications, or had homepages and publications that were identified by us. These assessments are depicted in the bottom left of Figure 100. Their MAP results looked similarly and the maximum MAP result of the CBCF baseline models was found in the CBCF SVD model with 700 latent topics.

As determined from the external validity analysis of Study 1, the bibliography-augmented 10-NN.PO full-text CN3-term CBF model and the bibliography-augmented SVD CN3-term CBCF model with 900 latent topics were the fusion models from the bibliography source. The external-bookmark-augmented centroid SVD CN3-term CBF model with 400 latent topics and the external-bookmark-augmented SVD CN3-term CBCF model with 1200 latent topics were the fusion model from the external bookmark source. The centroid SVD baseline model with 100 latent topics and the SVD CBCF baseline model with 700 latent topics were chosen as baselines for this evaluation.

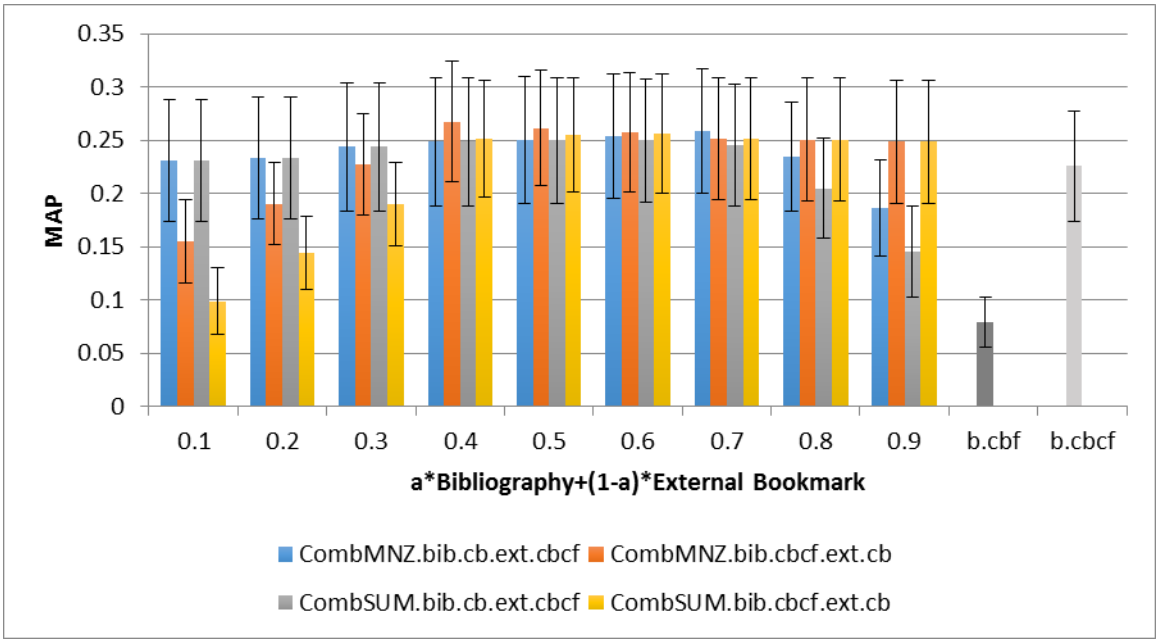


Figure 101: Bibliography + External Bookmark Fusion MAP of the External Validity

The experimental CombMNZ and CombSUM fusion approaches, which fused between the bibliography-augmented models and the external-bookmark-augmented ones, were assessed with 15 users who provided their publications and external scholarly bookmark accounts, and who bookmarked talks in any of the six holdout conferences, as shown in Figure **101**.

One-way ANOVA with Tuky's HSD Post Hoc test was applied to test the MAP results. The MAP results of experimental bibliography-augmented and external-bookmark-augmented recommendation fusion models varied from 0.1 to 0.27, depending on the type of fusion, recommending method, and the weight of fusion. Models fusing with CombMNZ method were quite sensitive to stepping weights. The poorest MAP results of the CombMNZ models were 0.15 at the fusion of the bibliography-augmented CBCF and the external-bookmark-augmented CBF models at a weight of 0.1, and 0.19 of the bibliography-augmented CBF and the external-

bookmark-augmented CBCF models at a weight of 0.9. However, the maximum MAP result of the CombMNZ fusion of the bibliography-augmented CBCF and the external-bookmark-augmented CBF models at a weight of 0.4 was 0.27. The peak MAP result of the CombMNZ fusion of the bibliography-augmented CBF and the external-bookmark-augmented CBCF models at a weight of 0.7 was 0.26. The CombSUM models were quite sensitive to stepping weights. The poorest MAP results of the CombSUM models were 0.1 at the fusion of the bibliography-augmented CBCF and the external-bookmark-augmented CBF models at a weight of 0.1, and 0.15 of the bibliography-augmented CBF and the external-bookmark-augmented CBCF models at a weight of 0.9. However, the maximum MAP result of the CombSUM fusion of the bibliography-augmented CBCF and the external-bookmark-augmented CBF models at a weight of 0.6 was 0.26. The peak MAP result of the CombSUM fusion of the bibliography-augmented CBF and the external-bookmark-augmented CBCF models at a weight of 0.6 was 0.25.

In summary, as shown in from Figure **101**, 32 out of 36 fusion models (17 CombMNZ models and 15 CombSUM models) outperformed significantly compared to the CBF baseline at a p-value < 0.05 level of significance. However, none of them performed significantly better than the CBCF baseline significantly.

9.4.2 Same-Source Fusion

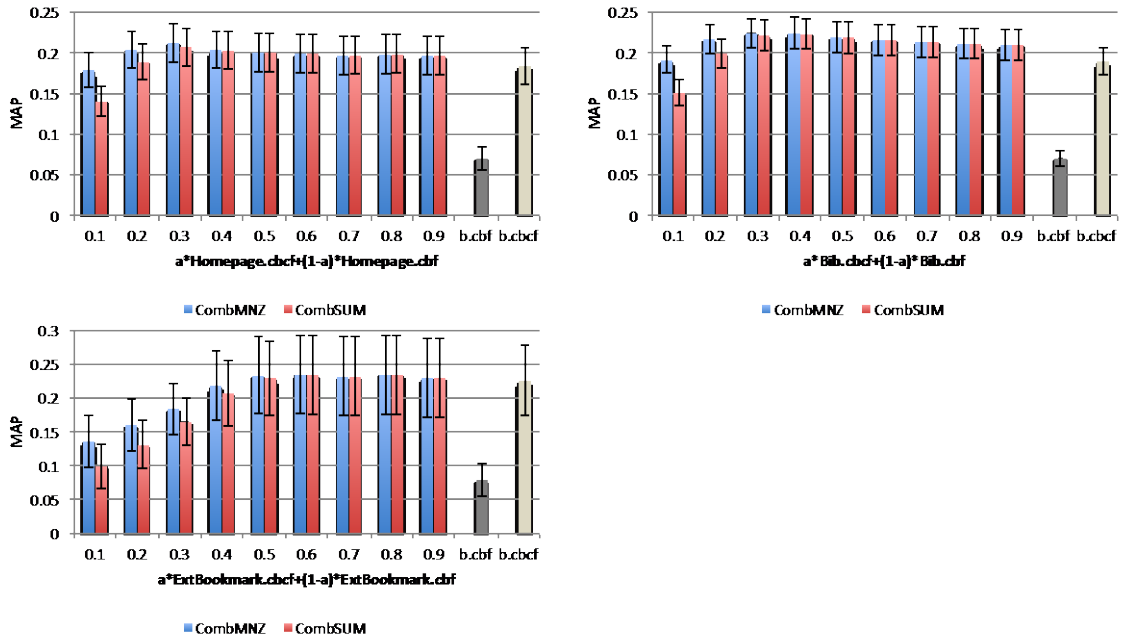


Figure 102: Same-Source Fusion Models of the External Validity

9.4.2.1 Homepage

The recommendations from the two homepage augmentation models, which were the homepage-augmented 35-NN.PO SVD CN3-term CBF model with 200 latent topics and the homepage-augmented SVD all-term CBCF model with 1000 latent topics, were combined and evaluated against two baselines, which were the 15-NN.PO SVD baseline CBF model with 100 latent topics and the SVD baseline CBCF model with 700 latent topics. These four models were assessed with 59 users who provided their homepage information or had their homepage information identified by us, and who also bookmarked talks in any of the six holdout conferences.

One-way ANOVA with Tuky's HSD Post Hoc test was applied to test the MAP results. The MAP results of experimental same-source-different-method homepage-augmented recommendation fusion models varied from 0.14 to 0.21, depending on the type of fusion, recommending method, and the weight of fusion. Models fusing with the CombMNZ method performed well on any stepping weights. The range of MAP results of CombMNZ models ranged between 0.18 and 0.21. The CombSUM models also performed well on any stepping weights. The maximum MAP result of the CombSUM fusion model at a weight of 0.3 was 0.21.

In summary, from the top left of Figure **102**, none of 18 homepage-augmented fusion models performed significantly better than the CBCF baseline significantly. However, all of them outperformed the CBF baseline at a p-value < 0.05 level of significance.

9.4.2.2 Bibliography

The recommendations from the two bibliography augmentation models, which were the bibliography-augmented 10-NN.PO full-text CN3-term CBF model and the bibliography-augmented SVD CN3-term CBCF model with 1000 latent topics, were combined and evaluated against two baselines, which were the 10-NN.PO SVD baseline CBF model with 200 latent topics and the SVD baseline CBCF model with 700 latent topics. These four models were assessed with 117 users, who provided their publications or had their publications identified by us, and who also bookmarked talks in any of the six holdout conferences.

One-way ANOVA with Tuky's HSD Post Hoc test was applied to test the MAP results. The MAP results of experimental same-source-different-method bibliography-augmented recommendation fusion models varied from 0.15 to 0.22, depending on the type of fusion, recommending method, and the weight of fusion. Models fusing with the CombMNZ method

performed well on any stepping weights. The maximum MAP result of the CombMNZ fusion bibliography-augmented models at a weight of 0.3 was 0.22. The CombSUM models also performed well on any stepping weights. The maximum MAP result of the CombSUM fusion model at a weight of 0.4 was 0.22.

In short, from the top right of Figure **102**, nine of 18 (five CombMNZ and four CombSUM models) homepage-augmented fusion models performed significantly better than the CBCF baseline at a p-value < 0.05 level of significance. All of them also outperformed the CBF baseline at a p-value < 0.05 level of significance.

9.4.2.3 Bookmarked Scholarly Papers (External Bookmark)

The recommendations from two external bookmark augmentation models, which were the external-bookmark-augmented centroid SVD CN3-term CBF model with 400 latent topics and the external-bookmark-augmented SVD CN3-term CBCF model with 1200 latent topics, were combined and evaluated against two baselines, which were the centroid SVD baseline CBF model with 100 latent topics and the SVD baseline CBCF model with 700 latent topics. These four models were assessed with 15 CN3 users who provided their external bookmark accounts or had their accounts identified by us, and who also bookmarked talks in the six holdout conferences.

One-way ANOVA with Tuky's HSD Post Hoc test was applied to test the MAP results. The MAP results of experimental same-source-different-method external bookmark-augmented recommendation fusion models varied from 0.1 to 0.23, depending on the type of fusion, recommending method, and the weight of fusion. Models fusing with the CombMNZ method performed in a way that suggested they were quite sensitive to stepping weights. The poorest

MAP result of CombMNZ models was 0.14 at a stepping weight of 0.1. However, the maximum MAP result of the CombMNZ fusion external bookmark-augmented models at a weight of 0.5 was 0.23. The CombSUM models also were also quite sensitive to stepping weights. The poorest MAP result of CombSUM models was 0.1 at a stepping weight of 0.1. However, the maximum MAP result of the CombSUM fusion model at a weight of 0.6 was 0.23.

Conclusively, from the bottom left of Figure 102, 15 of 18 external-bookmark-augmented fusion models (eight CombMNZ and seven CombSUM models) outperformed the CBF baseline at a p-value < 0.05 level of significance. However, none of them significantly outperformed the CBCF baseline.

9.5 SUMMARY AND DISCUSSION

In this chapter, six external-source-augmented models were selected from refitted models with 807 talks from six holdout conferences, in order to be reassessed with 122 users to validate the findings in Chapter 6 through Chapter 8. The validation processes were repeated for all three CN3 studies, which were study 1 – external source augmentation improvement; Study 2 – the cold-start problem; and Study 3 – the recommendation fusion. The external validity was re-evaluated with 122 qualified users.

In the reevaluation on Study 1, external source augmentation improvement, four models augmented with homepage and external bookmark sources were tested, and it was confirmed that they did not improve the performance from the baselines used in Study 1 in Chapter 6. Furthermore, the bibliography-augmented CBCF model did not improve their performance

relative to the baselines in Study 1. As expected, the bibliography-augmented CBF model improved the performance of the baseline significantly.

In the reevaluation on Study 2, the cold-start problem, homepage, bibliography, and external-bookmark augmentation models helped improve recommendation results for users had no bookmarked talks significantly, just as they did the cold-start problem study in Chapter 7. In detail, the homepage-augmented CBF model performed significantly better than the CBF baseline until the analysis was applied to situations that users had two bookmarked talks. The bibliography-augmented CBF model of the external validity of the cold-start problem study was able to improve the performance of baseline until there were eight bookmarked talks in the user profiles; it almost accomplished the ten-bookmarked-talks milestone discussed in section 7.3.2.
In the last study, recommendation fusion, there were 151 out of 162 (93.20%) different-approach CombMNZ or CombSUM recommendation fusion models, comprising 100 cross-source and 51 same-source models. These models significantly outperformed the CBF baseline models. However, when comparing them with CBCF baseline models, only 11 out of 162 (6.79%) different-approach CombMNZ or CombSUM recommendation fusion models (which consisted of two cross-source and nine same-source models) significantly improved the performance of baseline recommendations. The recommendation fusion results confirmed the external validity of the fusion models from Study 3.

10.0 STUDY 4: EVALUATING MODELS IN A USER STUDY

This chapter explains the user study on the CoMeT system. First, it describes the CoMeT research talks and how they were used in the experiment. Second, it describes how participant demographics and their external source information, such as personal websites, bibliographies, and external bookmarks, were used in the experiment. Third, it describes the procedure for this user study. Fourth, it reviews the consistency of the subjects' relevance and novelty judgments. Fifth, the results of each experiment are reviewed. Lastly, the chapter draws conclusions and discusses of the results.

10.1 COMET DATA

The research talks were selected from the CoMeT system and divided into two sets, a training set and a test set. The training set consisted of 271 talks, starting in the week of September 10, 2012 and ending in the week of November 2, 2012. The test set was comprised of 336 talks, which occurred from the week of February 4, 2013 to the week of March 29, 2013. After removing stop words and stemming data, there were 8975 unique unigram terms, consisting of 5401 terms in the training set and 6922 terms in the test set, respectively. There were 3348 overlapping terms between the training and the test sets. Figure **103** shows the 100 terms that appeared most

frequently in the dataset. Figure **104** and Figure **105** go into further detail by showing the top 100 terms in the training and the test sets.

	Term	Number of Talks		Term	Number of Talks
1	research	229	49	network	79
2	talk	220	49	behavior	79
3	information	209	53	large	78
4	university	208	53	engineering	78
5	work	195	55	institute	76
6	model	193	55	challenge	76
7	data	167	55	project	76
8	system	166	58	school	75
9	study	162	58	department	75
10	abstract	159	58	technique	75
11	result	155	61	include	74
11	contact	155	61	host	74
13	science	153	63	interaction	73
14	present	145	64	computation	72
15	discuss	140	64	people	72
16	show	136	64	propose	72
17	sponsor	134	64	group	72
18	focus	132	68	important	71
19	professor	128	69	understanding	70
20	problem	121	69	make	70
21	including	119	71	role	69
22	design	116	71	improve	69
22	process	116	73	demonstrate	68
24	develop	115	73	structure	68
25	time	114	73	case	68
25	technology	114	76	field	67
25	base	114	76	degree	67
28	computer	105	78	machine	66
29	recent	104	78	number	66
30	learning	103	80	investigate	65
31	application	101	80	2012	65
32	approach	98	80	individual	65
33	theory	96	83	member	64
34	development	95	84	compute	63
35	faculty	92	84	national	63
36	years	91	86	form	62
37	describe	90	86	award	62
38	social	89	86	world	62
39	algorithm	86	86	explore	62
39	provide	86	90	framework	61
39	received	86	90	issue	61
42	center	84	92	current	60
42	human	84	92	function	60
44	tepper	83	94	examine	59
44	method	83	94	environment	59
46	program	81	96	language	58
47	analysis	80	96	experiment	58
47	paper	80	96	find	58
49	serve	79	99	identify	57
49	support	79	99	analyze	57

Figure 103: The Top 100 Terms in the Dataset with their Document Frequency

	Term	Number of Talks		Term	Number of Talks
1	research	106	49	institute	36
2	information	98	52	program	35
3	talk	97	52	department	35
4	model	91	52	method	35
5	university	89	52	group	35
6	work	80	52	computation	35
7	data	79	57	technique	34
8	study	70	57	people	34
9	science	69	57	make	34
10	system	68	60	understanding	33
11	result	67	60	framework	33
12	show	65	60	network	33
12	sponsor	65	60	include	33
14	contact	63	60	support	33
14	present	63	65	explore	32
16	base	62	65	language	32
17	focus	61	65	field	32
17	discuss	61	65	demonstrate	32
17	including	61	65	project	32
20	professor	60	70	degree	31
21	problem	59	71	national	30
22	technology	56	71	compare	30
23	process	52	71	significant	30
23	develop	52	71	center	30
25	application	51	71	mellon	30
26	approach	50	71	tool	30
27	abstract	48	71	serve	30
28	computer	47	71	area	30
29	learning	46	71	evaluate	30
29	design	46	71	structure	30
31	time	44	71	compute	30
32	describe	43	82	committee	29
32	recent	43	82	author	29
34	social	42	82	propose	29
34	years	42	82	identify	29
34	provide	42	82	school	29
37	development	41	82	online	29
37	behavior	41	82	machine	29
39	2012	40	89	host	28
40	algorithm	39	89	context	28
41	analysis	38	89	found	28
42	large	37	89	finding	28
42	faculty	37	89	director	28
42	theory	37	89	interaction	28
42	human	37	95	general	27
42	tepper	37	95	user	27
42	received	37	95	carnegie	27
42	paper	37	95	fellow	27
49	case	36	95	analyze	27
49	engineering	36	95	improve	27

Figure 104: The Top 100 Terms in the Training Set with their Document Frequency

	Term	Number of Talks		Term	Number of Talks
1	talk	123	46	school	46
2	research	123	52	important	45
3	university	119	52	interaction	45
4	work	115	52	role	45
5	abstract	111	52	form	45
5	information	111	56	individual	44
7	model	102	56	provide	44
8	system	98	56	project	44
9	contact	92	59	propose	43
10	study	92	59	paper	43
11	data	88	61	analysis	42
11	result	88	61	engineering	42
13	science	84	61	improve	42
14	present	82	64	large	41
15	discuss	79	64	technique	41
16	focus	71	64	include	41
16	show	71	67	institute	40
18	time	70	67	department	40
18	design	70	69	examine	39
20	sponsor	69	69	number	39
21	professor	68	71	behavior	38
22	process	64	71	structure	38
23	develop	63	71	world	38
24	problem	62	71	investigate	38
25	recent	61	71	people	38
26	theory	59	76	machine	37
27	technology	58	76	computation	37
27	computer	58	76	member	37
27	including	58	76	group	37
30	learning	57	76	award	37
31	faculty	55	76	understanding	37
32	development	54	82	degree	36
33	center	54	82	demonstrate	36
34	base	52	82	part	36
35	application	50	82	issue	36
36	serve	49	82	make	36
36	challenge	49	87	current	35
36	received	49	87	field	35
36	years	49	87	function	35
40	approach	48	90	experiment	34
40	method	48	90	address	34
42	describe	47	92	control	33
42	algorithm	47	92	potential	33
42	social	47	92	national	33
42	human	47	92	mechanism	33
46	network	46	92	scale	33
46	program	46	92	compute	33
46	tepper	46	92	evidence	33
46	host	46	92	suggest	33
46	support	46	92	experience	33

Figure 105: The Top 100 Terms in the Test Set with their Document Frequency

10.2 EXTERNAL DATA AND PARTICIPANT DEMOGRAPHICS

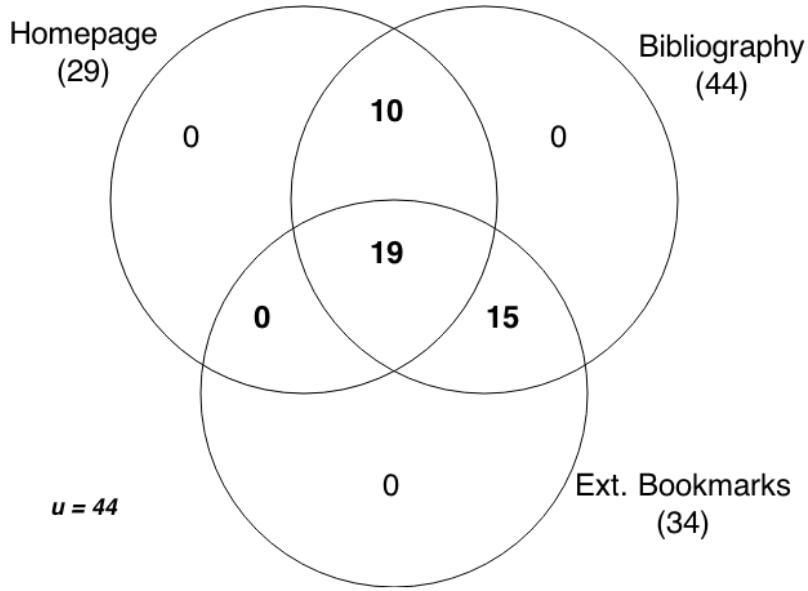


Figure 106: Participants' Demography in the CoMeT User Study

Forty-four graduate students from the University of Pittsburgh and Carnegie Mellon University were recruited as subjects for the experiment. One set of experiment results was dropped because the number of bookmarked talks in the training set was insufficient to conduct the five-fold cross validation. This type of analysis required that each subject have at least five bookmarked talks in the training set. The participants were individuals who were publishing or have published papers in academic conferences. They provided an account of at least one external source, homepage, or social bookmarking systems, including CiteULike, and Mendeley. Figure **106** shows participants' demographic information, including whether they had a personal webpage (29 subjects), a bibliography (44 subjects), and/or external scholarly bookmark articles (44 subjects).

In this user study, on average, subjects bookmarked more than just the off-line CN3 validation. Out of the 44 subjects, 37 of them (84.09%) bookmarked more than 20 CoMeT talks in the training set, as shown in Figure **107**. The proportion of participations with a high number of bookmarks remains the same (84.09%) in the test set, as shown in Figure **108**.

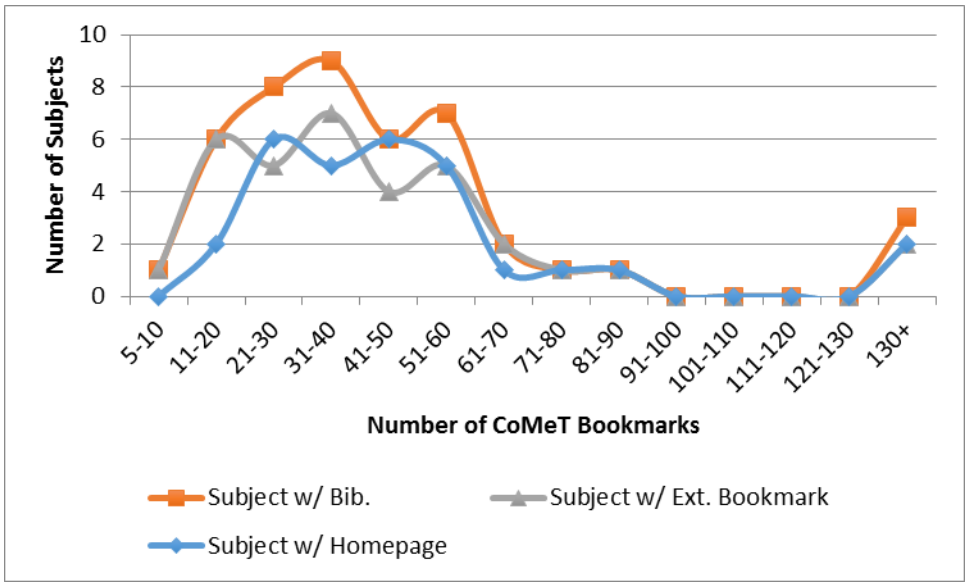


Figure 107: CoMeT Bookmark Distributions in the Training Set

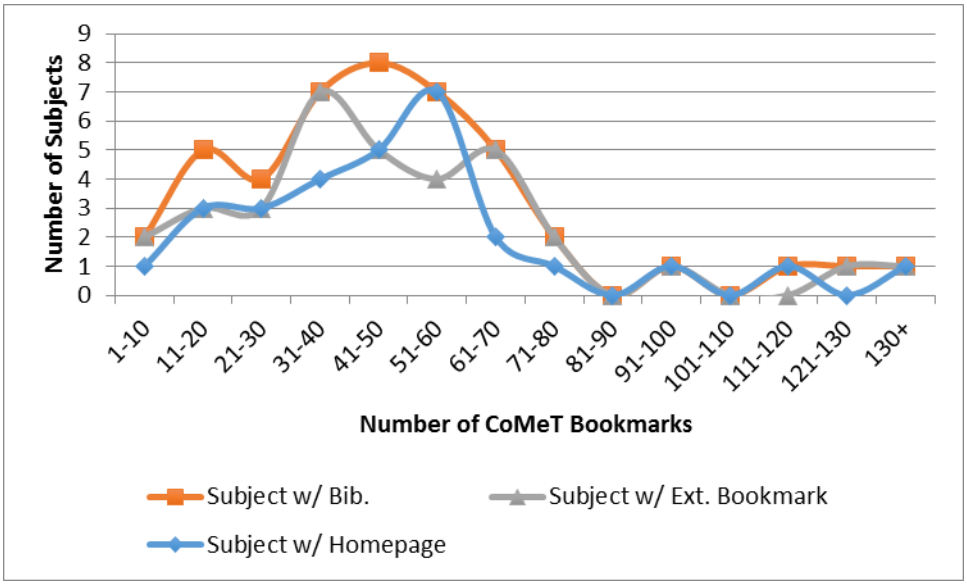


Figure 108: CoMeT Bookmark Distributions in the Test Set

Figure **109** shows a distribution of the number of pages in the subjects' homepages. Most of the subjects' homepages (72.41%) contained one to five pages. The total number of pages from all the participants combined was 126. A number of participants also reported a low number of publications; 63.63% of them had published 10 or fewer articles, as shown in Figure **110**. The total number of publications from all subjects combined was 418. Based on the the data from 34 participants who provided their external bookmarked papers, 64.71% of their bookmarked articles had 50 or fewer papers, as shown in Figure 111. The total number of external bookmarked articles from all participants combined was 2276.

Table 4: The Number of Total Unigram Terms in the External Sources

	#Total Terms	#Terms overlapped with CoMeT	#Terms after combining with CoMeT
Homepage	4928	2668	11235
Bibliography	3532	2513	9994
External Bookmark	8744	4315	13404

Looking into unigram terms of external sources, these sources introduced extra terms into consideration. Table 4 shows a summary of the unigram terms in the external sources. There were 2260 extra terms increased from the homepage, 1019 extra terms increased from the bibliography, and 4429 extra terms from the external scholarly bookmarked scholarly articles sources.

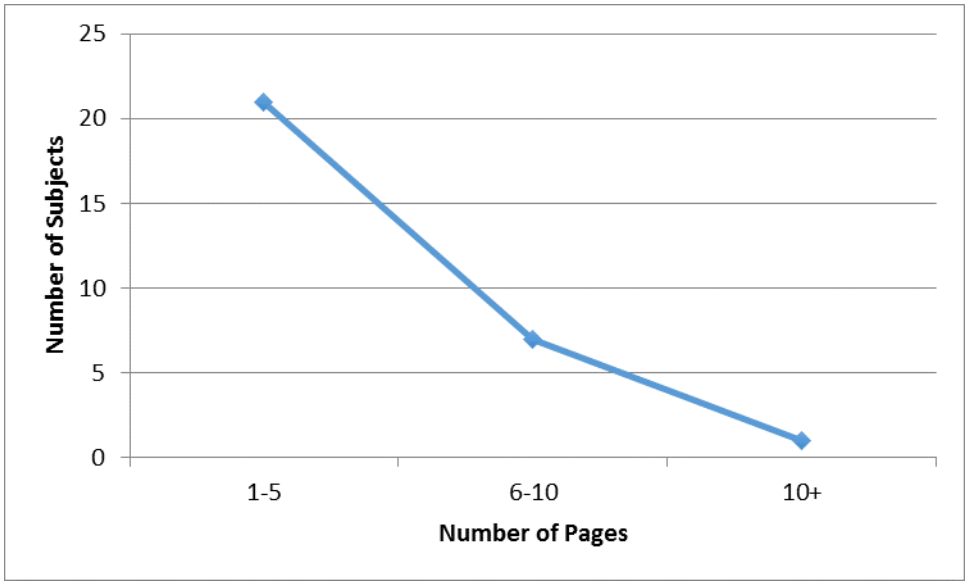


Figure 109: Homepage Distribution in the CoMeT User Study

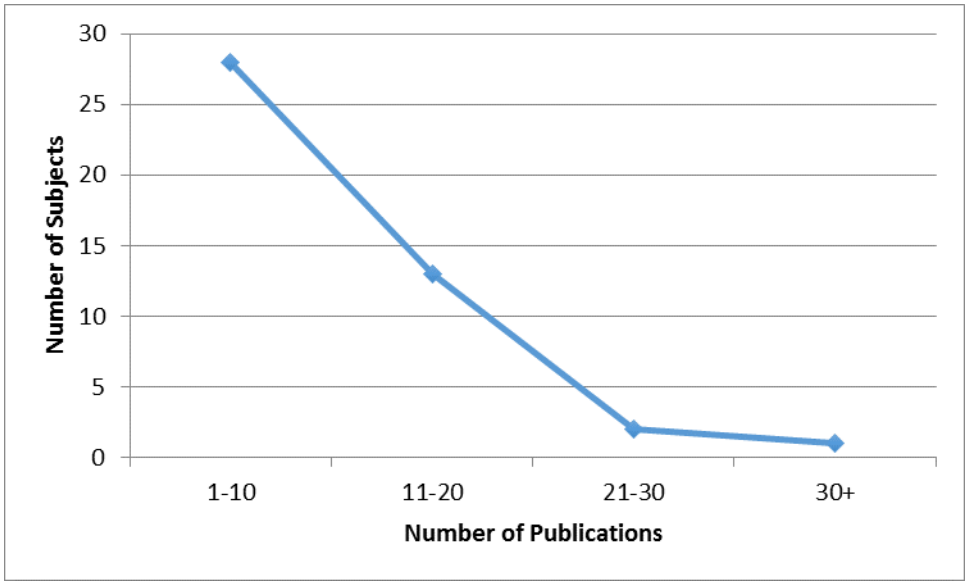


Figure 110: Bibliography Distribution in the CoMeT User Study

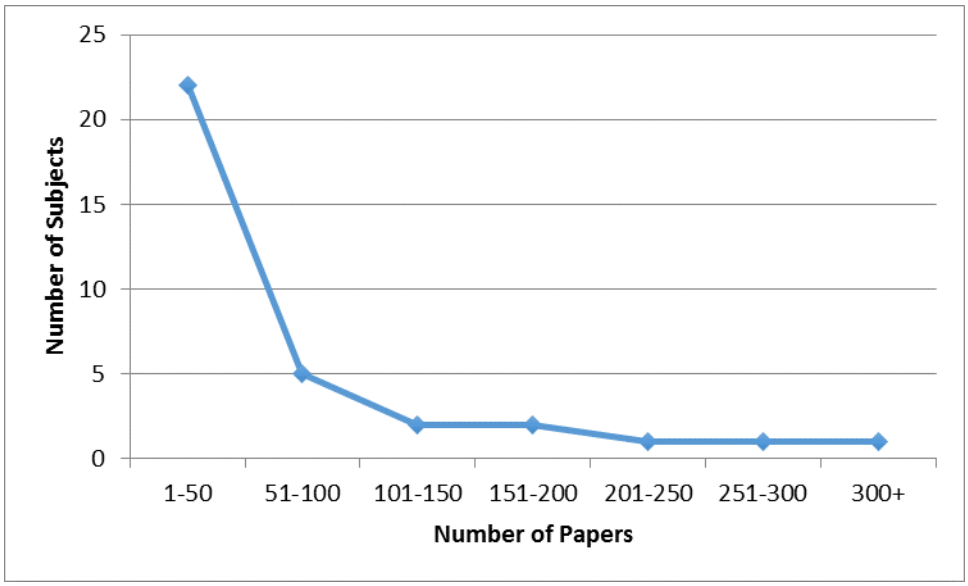


Figure 111: External Scholarly Bookmarked Papers Distribution in the CoMeT User Study

Table 5: A Summary of the Participants' Demographic Information

		Subject	Homepage	Bibliography	External Bookmark
Gender	Female	17	13	17	13
	Male	27	16	27	21
Age	20-25	9	8	9	6
	26-30	18	11	18	14
	31-35	14	10	14	11
	>= 36	3	-	3	3
Level of Study	Graduate (Master)	2	2	2	1
	Ph.D. Student	7	4	7	5
	Preliminary	7	5	7	5
	Comprehensive/Qualified	15	11	15	11
	Proposal	7	3	7	6
	Defense	4	3	4	4
	Post Doctoral	2	1	2	2
Major of Study	Computer Science	11	10	11	7
	Information Science	18	11	18	14
	Others	15	8	4	13

10.3 STUDY PROCEDURE

Objective. This experiment was conducted in order to answer the fourth research question, which required the user study to find the relevance and novelty effects of the recommendations from the previous three experiments.

Design. The tuning models' parameters and procedures needed to be repeated in order to fit models in the training set of the CoMeT dataset. Five-fold cross validation was used to

determine the candidate models, using both external sources and recommendation approaches (CBF or CBCF).

To remain consistent with the previous CN3 studies, and to reduce the complexity and loads of the study, the clustering approaches were dropped and the extra unigram terms were excluded, leaving only those appearing in the CoMeT corpus. The best-performing models, as determined by MAP measure, external source, and recommendation approach, were evaluated in the test set for Study 4.1 and Study 4.3. The evaluation assessed the external-source-augmented models against baselines on three measures: MAP, Relevance Performance, and Novelty Performance.

The 20 bookmarked talks in the training set were randomly selected for the cold-start problem study. Those talks were then randomly assigned to the cold-start windows. The cold-start window size was from zero to 20 bookmarked talks. For each cold-start window, the process was repeated 10 times. The evaluation results of each cold-start window were the average of the results of the 10 iterations.

Setting. In the first part of the study, subjects were given a brief five-minute training sessions to ensure that they knew how to bookmark talks in the system. Then, they were asked to bookmark the talks that they considered attending within the whole training set. In the second part of the study, subjects were given another brief five-minute training session to make sure that they understood how to rate talks in the system. They were asked to bookmark talks that they would consider attending within the whole test set. If they bookmarked a talk, then they had to give their ratings of its relevance, and novelty in relation to their interest in the test set, using a five-level Likert scale. The stepping scale allowed participants to rate their answers on a scale of 1 to 5, where ‘1’ means not relevant at all/not novel at all, ‘2’ means probably not

relevant/probably not novel, ‘3’ means neutral, ‘4’ means probably relevant/probably novel, and ‘5’ means extremely relevant/extremely novel. The subjects in this experiment were considered experts. Their ratings were considered perfect.

The two ratings, relevance and novelty, were used to rank each bookmarked talk for the perfect order of the test set, which was then used as the normalization constant for nDCG. These two perfect orders also were used to compute the nDCG of the results by all other recommendation methods.

The research interest: relevance and research interest: novelty ratings of each talk in the test set were recorded. The value of nDCG for each approach in the two judgments will be calculated. Fleiss’ kappa was used to measure the consistency of each of the subjects’ two judgments (relevance, and novelty). One-way ANOVA was applied to test each of the three hypotheses. The null hypothesis was rejected if the results from the F-test indicated a significant difference, using a p-value < 0.05 standard of significance. when the null hypothesis was rejected, all pairwise differences were examined with the Scheffe procedure.

Dependent Variables. The dependent variables were nDCG, and MAP results. The recommendations returned from six experimental approaches with external sources augmentation were expected to yield higher MAP results and higher NDCG results than the recommendation results that were returned from the no-external-source baselines.

Hypotheses of Study 4

Research Question 4.

“*In a realistic recommendation context, how do external-source-augmentation recommendations affect the accuracy, relevance, and novelty of recommended talks?*”

Research Question 4A Accuracy Performance

H_0: There is no statistically significant difference between the *accuracy* means of the recommended talks and those from the *baseline, CBF with external source augmentation, CBCF with external source augmentation,* and *fusion with external source augmentation*.

Metrics: ***MAP***

Research Question 4B Relevance Performance

H_0: There is no statistically significant difference between the *relevance* means of the recommended talks and those from the *baseline, CBF with external source augmentation, CBCF with external source augmentation,* and *fusion with external source augmentation*.

Metrics: ***NDCG***

Research Question 4C Novelty Performance

H_0: There is no statistically significant difference between the *novelty* means of recommended talks and those from the *baseline, CBF with external source augmentation, CBCF with external source augmentation,* and *fusion with external source augmentation*.

Metrics: ***NDCG***

10.4 THE CONSISTENCY OF THE SUBJECTS' JUDGMENTS

Fleiss' kappa was selected as a means to measure the consistency of the subjects' relevance and novelty judgments. The Fleiss' kappa is a statistical measure for assessing the reliability of agreement between raters who assign or classify a fixed number of items. It shows the degree of agreement that would be expected by chance. The equations for Fleiss' Kappa are shown below:

$$k = \frac{\overline{P} - \overline{P}_e}{1 - \overline{P}_e}$$

$$\overline{P} = \frac{1}{N}\sum_{i=1}^{N} p_i$$

$$\overline{P}_e = \sum_{j=1}^{k} p_j^2$$

$$p_i = \frac{1}{m(m-1)}\left(\sum_{j=1}^{k} x_{ij}^2 - m\right)$$

$$p_j = \frac{1}{Nm}\sum_{i=1}^{N} x_{ij}$$

Where,

k is the kappa,

$\overline{P}$ is the mean of the extent to which the rater agrees for the item p_i,

$\overline{P}_e$ is the sum of square of the proportion of all assignments to the category p_j,

N is the total number of items,

m is the number of ratings per item,

k is the number of categories, and

x_{ij} is the number of raters who assigned the item p_i to the category p_j.

The subjects were asked to bookmark talks in the training set in relation to their attendant interests. Later, subjects were asked to bookmark talks in the test set regarding to their attendant interests, and also rate their relevance and novelty in relation to their interests. As a result, there are four judgment classes of subjects to be tested on Fleiss' kappa: attending judgment in the training set, attending judgment in the test set, relevance judgment in the test set, and novelty judgment in the test set. Every talk had the same number of raters for the attending judgment in both the training and the test sets. However, each talk had a different number of raters for the relevance and novelty judgments.

Figure **112** shows the number of subjects who bookmarked each talk in the training set, and Figure **113** shows the number of subjects who bookmarked each talk in the test set. As shown in the figures, there were four talks in the training set and there were 29 talks in the test set that were not bookmarked at all. The subjects only made relevance and novelty judgments for talks that they bookmarked in the test set. In order to calculate Fleiss' kappa, there must be at least two subjects assigning their ratings to certain items. As a result, 63 talks (18.75%) in the test set were dropped out of the Fleiss' Kappa assessment of relevance and novelty judgments, because they were only bookmarked by one subject or were not bookmarked at all.

Because all 44 subjects had to choose whether or not to bookmark in every talk in both the training and the test sets, the Fleiss' kappa is suitable to measure the consistency of their bookmarking judgments. In the Fleiss' kappa scale, $k = 1$ means the ratings are in the perfect agreement.

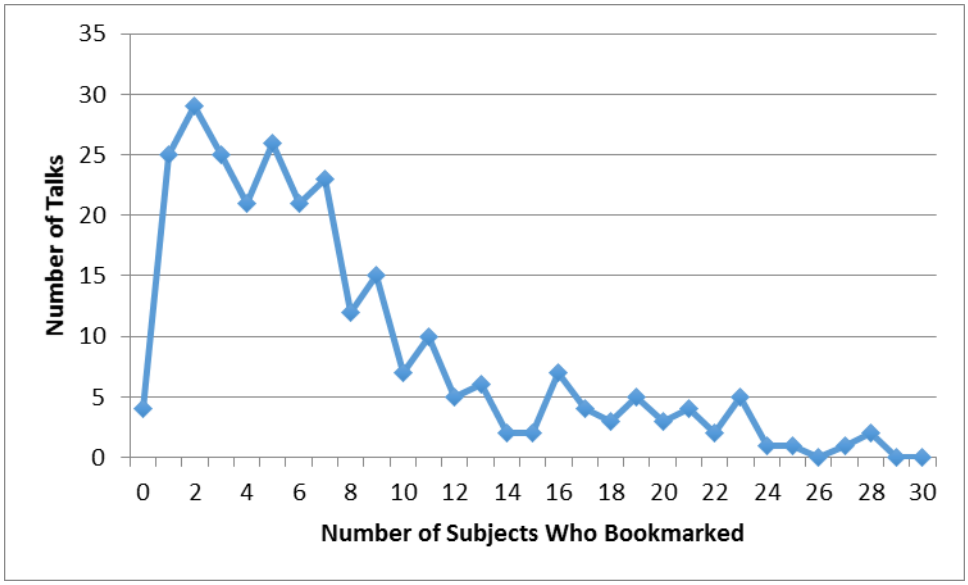


Figure 112: Number of Subjects who Bookmarked Each Talk in the Training Set

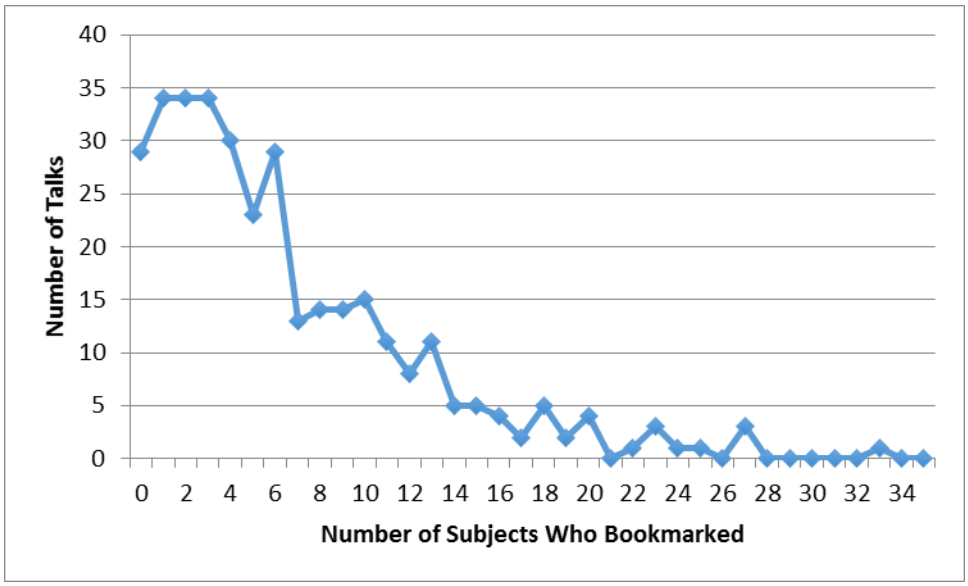


Figure 113: Number of Subjects who Bookmarked Each Talk in the Test Set

As shown in Table **6**, the subjects had low agreements about which talks they bookmarked in both the training and the test sets. In the other words, subjects we recruited had a diversity of

interests across the CoMeT talks. For talks that had at least two subject relevance judgments, on average, subjects had fair agreements on which talks they bookmarked. They had more agreements on the novelty judgments as well.

Table 6: The Means of the Extent to Which Subjects Agree About CoMeT Talks for Each Type of Judgment

Subject Judgment	Fleiss' Kappa
Bookmarking on Training Set	0.117869497
Bookmarking on Test Set	0.121076946
Relevancy	0.213992695
Novelty	0.285763805

10.5 RESULTS

10.5.1 Study 4.1: External-Source Augmented Recommendation Improvement

The first study aimed to investigate the impact of external source augmentation on recommendations. In this study, there were three external sources: homepage, bibliography, and external scholarly papers (external bookmarks). There were two main recommendation

approaches used in this study: content-based recommendation and content-boosted collaborative filtering.

10.5.1.1 Homepage

1) Training Set: Fitting Models

The CBF baselines for the 29 subjects who provided homepages were assessed in the training set, as shown in Figure **114**. The maximum MAP result of the individual centroid models was 0.21 on the full-text centroid model, which was the peak MAP result in comparison to other SVD centroid models, as shown in the top left diagram in Figure **114**. The SVD centroid model with 100 latent topics performed better than the other SVD models. Therefore, 100 latent topics were used later in the other SVD CBF baseline models. At the bottom left of Figure **114**, all the KNN.PO baseline models performed similarly. The full-text KNN.PO models performed better than the SVD models, but there was no significant difference among them. The maximum MAP result of the KNN.PO model was 0.22 for the 20-NN.PO full-text model.

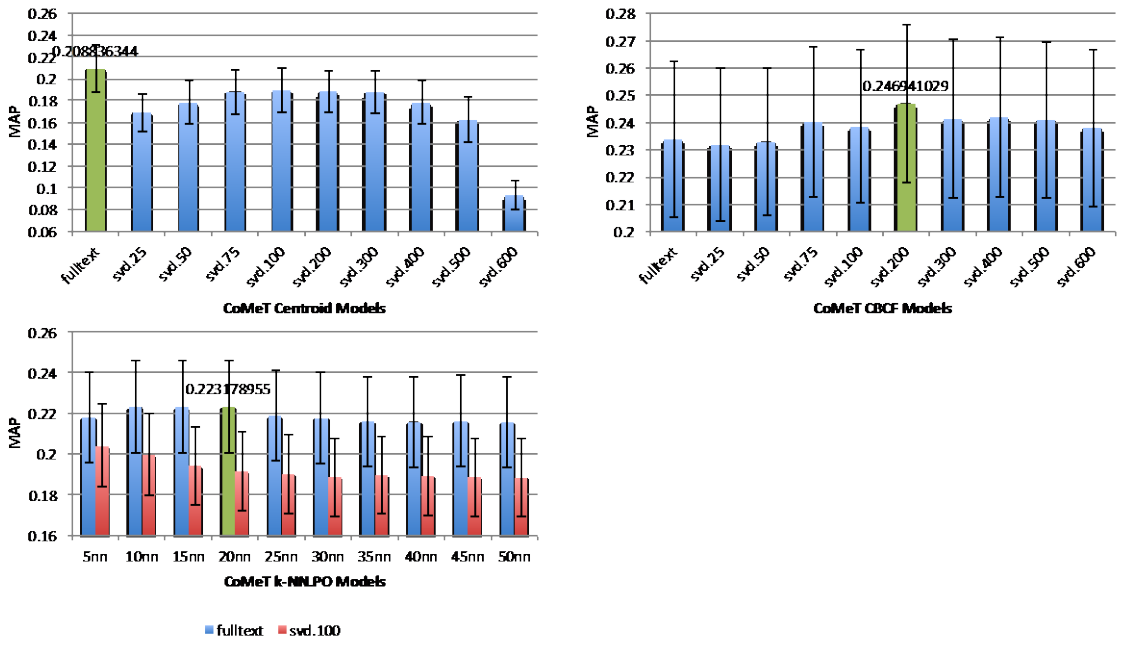


Figure 114: Homepage User Baseline Models

The CBCF baseline models were assessed with the 29 subjects who provided homepages in the training set, as depicted in the top right of Figure **114**. Their MAP results looked steadily, and the maximum MAP result of the CBCF baseline models was found in the CBCF SVD model with 200 latent topics.

The 20-NN.PO full-text baseline model was selected as a content-based baseline model, as shown in the top left and bottom left of Figure **114**. The SVD content-boosted collaborative filtering baseline with 200 latent topics was selected as the CBCF baseline representative, as shown in the top right of Figure **114**.

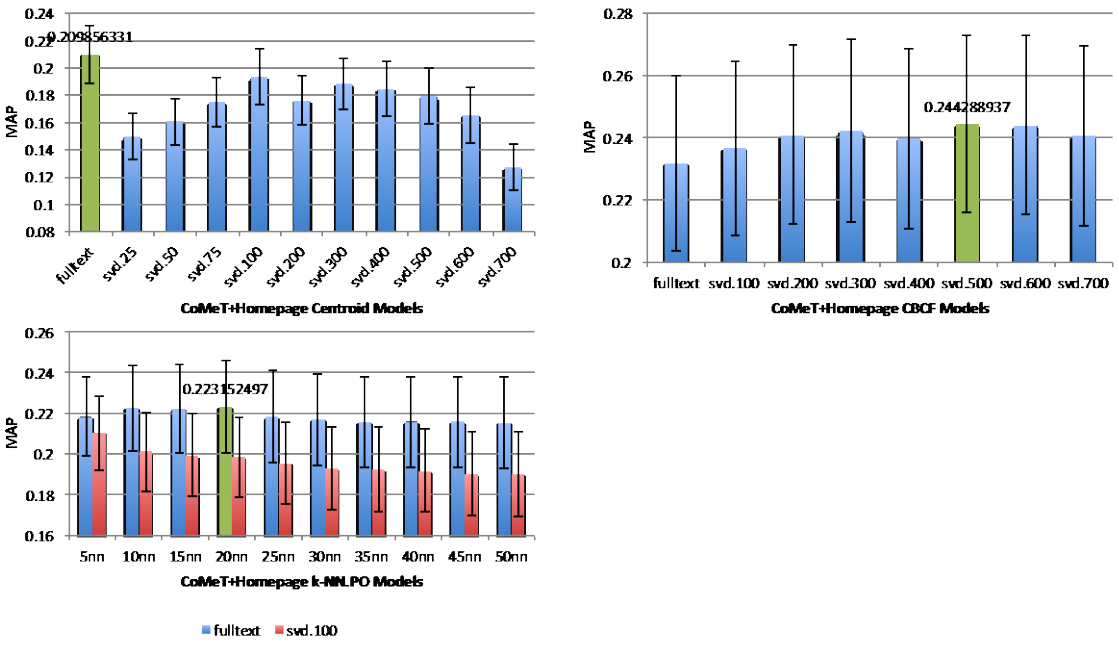
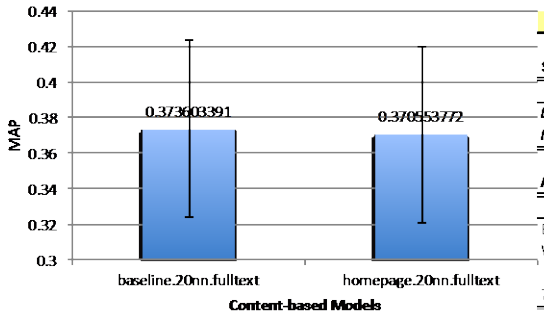


Figure 115: Homepage-Augmented Models

The homepage-augmented 20-NN.PO full-text model was chosen as the candidate for the homepage-augmented CBF model, because its mean MAP result was the highest among the other content-based models, as shown in the top left and bottom left of Figure **115**. The homepage-augmented SVD content-boosted collaborative filtering model with 500 latent topics, shown in the top right of Figure **115**, was selected as the representative of the homepage-augmented CBCF model, because its means MAP result was the highest among the other CBCF models.

2) Test Set: Evaluation

In the CBF comparison on MAP, both models were assessed with the 29 subjects who provided homepages in the test set. One-way ANOVA was applied to test the MAP results. The top right of Figure **116** shows that the homepage-augmented CBF model performed approximately the same level as the CBF baseline.



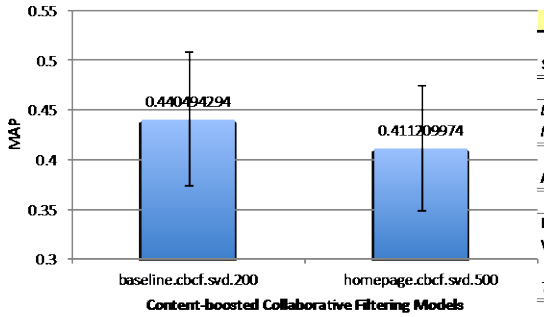
Analysis of Variance (One-Way)

Summary

Groups	Sample size	Sum	Mean	Variance
baseline.20nn.fulltext	29	10.8345	0.3736	0.01802
homepage.20nn.fulltext	29	10.74606	0.37055	0.01795

ANOVA

Source of Variation	SS	df	MS	F	p-level	F crit
Between Groups	0.00013	1	0.00013	0.0075	0.9313	5.73507
Within Groups	1.00714	56	0.01798			
Total	1.00727	57				



Analysis of Variance (One-Way)

Summary

Groups	Sample size	Sum	Mean	Variance
baseline.cbcf.svd.200	29	12.77433	0.44049	0.03255
homepage.cbcf.svd.500	29	11.92509	0.41121	0.0292

ANOVA

Source of Variation	SS	df	MS	F	p-level	F crit
Between Groups	0.01243	1	0.01243	0.40276	0.52825	5.73507
Within Groups	1.72893	56	0.03087			
Total	1.74136	57				

Figure 116: MAP Analysis of Homepage Models in the Test Set

In the CBCF comparison on MAP, the homepage-augmented SVD content-boosted collaborative filtering model with 500 latent topics was assessed with the same 29 subjects. As shown in the bottom right of Figure **116**, the homepage-augmented CBCF model performed lower than the CBCF baseline did, but the difference was not statistically different, using the p-value < 0.05 standard for significance.

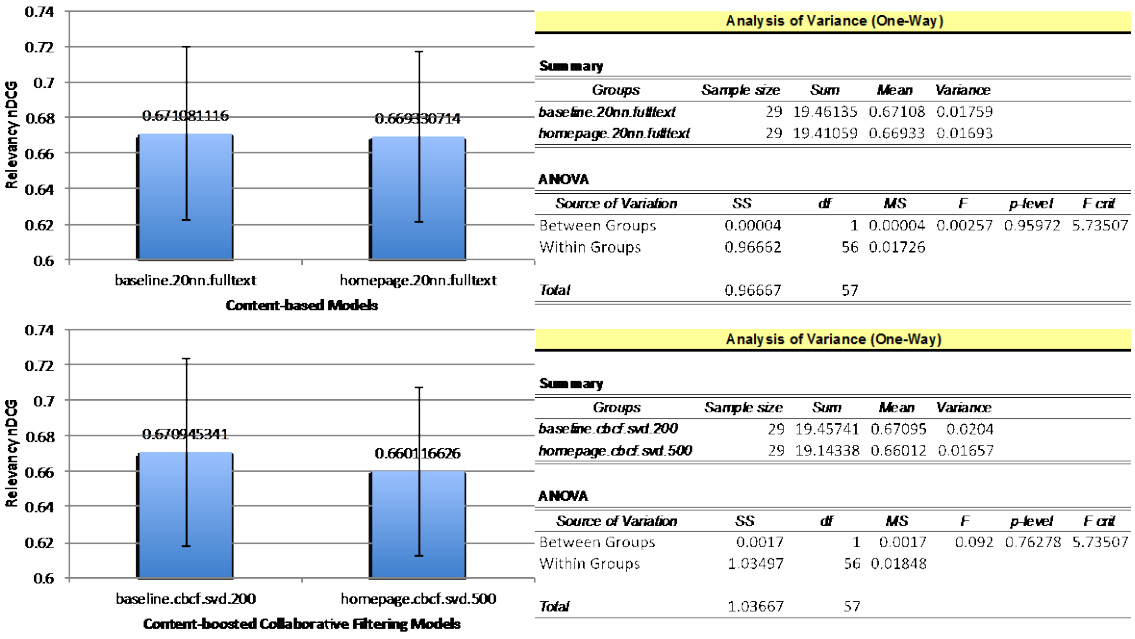
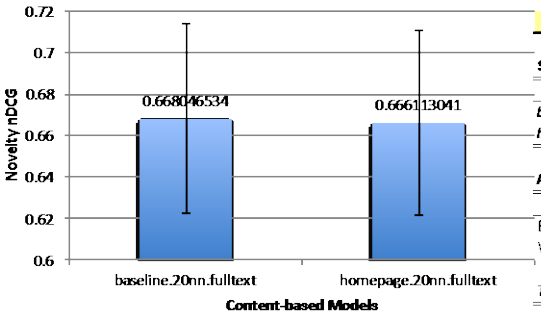


Figure 117: Relevance nDCG of Homepage Models on the Test Set

In the CBF comparison of Relevance nDCG, both models were assessed with the 29 subjects who provided homepages in the test set. One-way ANOVA was applied to test the Relevance nDCG results. As shown in the top right of Figure **117**, the homepage-augmented CBF model performed at approximately the same level as the CBF baseline.

In the CBCF comparison of Relevance nDCG, the homepage-augmented SVD content-boosted collaborative filtering model with 500 latent topics was assessed with the same 29 subjects. As shown in the bottom right of Figure **117**, the homepage-augmented CBCF model performed slightly lower than the CBCF baseline, but the difference was not statistically different, using the p-value < 0.05 level of significance.



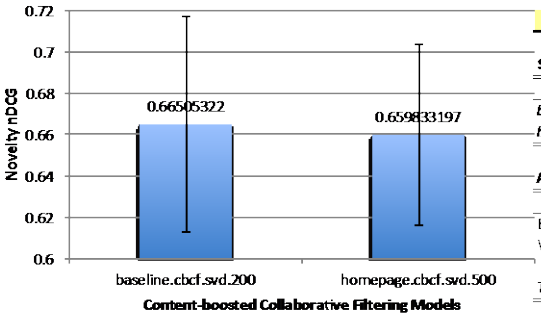
Analysis of Variance (One-Way)

Summary

Groups	Sample size	Sum	Mean	Variance
baseline.20nn.fulltext	29	19.37335	0.66805	0.0155
homepage.20nn.fulltext	29	19.31728	0.66611	0.01461

ANOVA

Source of Variation	SS	df	MS	F	p-level	F crit
Between Groups	0.00005	1	0.00005	0.0036	0.95236	5.73507
Within Groups	0.84301	56	0.01505			
Total	0.84307	57				



Analysis of Variance (One-Way)

Summary

Groups	Sample size	Sum	Mean	Variance
baseline.cbcf.svd.200	29	19.28654	0.66505	0.01969
homepage.cbcf.svd.500	29	19.13516	0.65983	0.01391

ANOVA

Source of Variation	SS	df	MS	F	p-level	F crit
Between Groups	0.0004	1	0.0004	0.02352	0.87866	5.73507
Within Groups	0.94072	56	0.0168			
Total	0.94112	57				

Figure 118: Novelty nDCG of Homepage Models in the Test Set

In the CBF comparison of Novelty nDCG, both models were assessed with the 29 subjects who provided homepages in the test set. One-way ANOVA was applied to test the Novelty nDCG results. As shown in the top right of Figure **118**, the homepage-augmented CBF model performed at approximately the same level as the CBF baseline.

In the CBCF comparison on Novelty nDCG, the homepage-augmented SVD content-boosted collaborative filtering model with 500 latent topics was assessed with the same 29 subjects. As depicted in the bottom right of Figure **118**, the homepage-augmented CBCF model performed at the same level as the CBCF baseline.

10.5.1.2 Bibliography

1) **Training Set: Fitting Models**

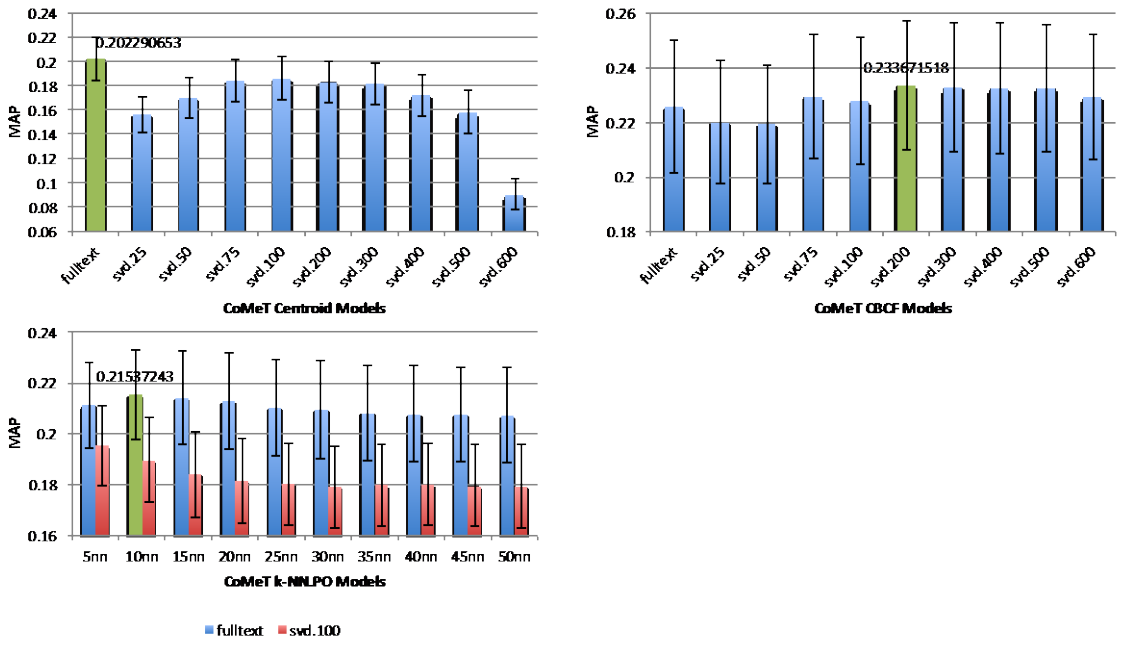


Figure 119: Bibliography-User Baseline Models

The CBF baselines were assessed with the 44 subjects who provided their publications in the training set, as shown in Figure **119**. The maximum MAP result of the individual centroid models was 0.20 on the full-text centroid model, which was the peak MAP result in comparison to the other SVD centroid models, as shown in the top left diagram on Figure **119**. The SVD centroid model with 100 latent topics also performed better than the other SVD models. As a result, 100 latent topics were used later in the other SVD CBF baseline models. At the bottom left of Figure **119**, all the KNN.PO baseline models performed steadily. The KNN.PO full-text models performed better than the SVD models, but there was no significant difference among them. The maximum MAP result of the KNN.PO model was 0.215 for the 10-NN.PO full-text model.

The CBCF baseline models were assessed with the same 44 subjects who provided their publications in the training set, as depicted in the top right of Figure **119**. Their MAP results

looked similarly and the maximum MAP result of CBCF baseline models was in the CBCF SVD model with 200 latent topics.

The 10-NN.PO full-text baseline model was selected from the others as a content-based baseline model, as shown in the top left and bottom left of Figure **119**. The SVD content-boosted collaborative filtering baseline with 200 latent topics was also selected as the CBCF baseline representative, as shown in the top right of Figure **119**.

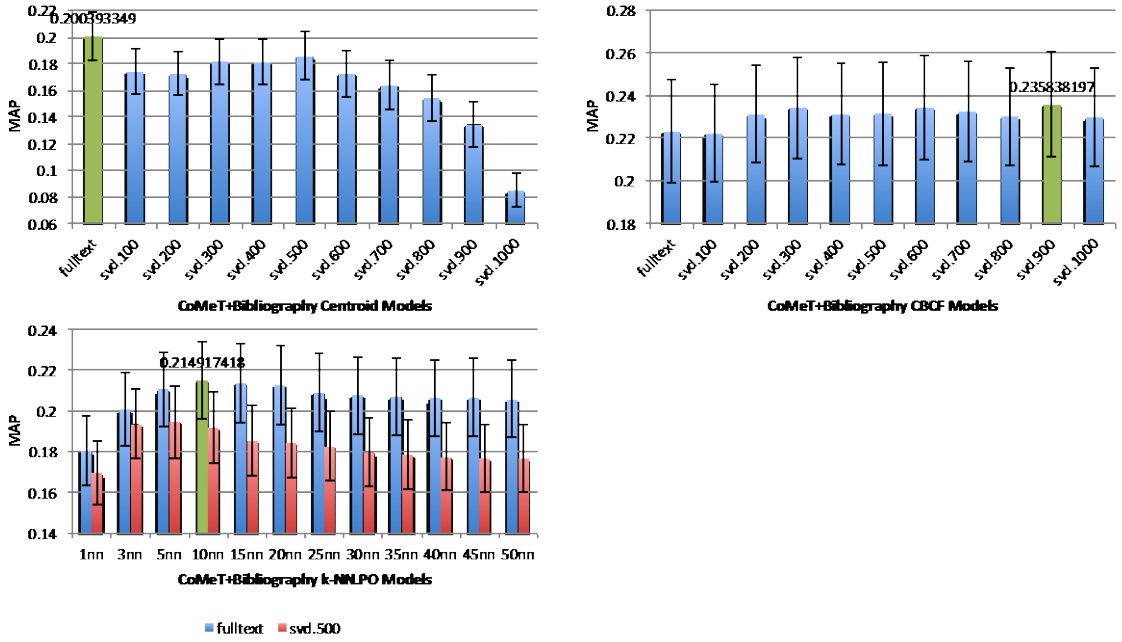
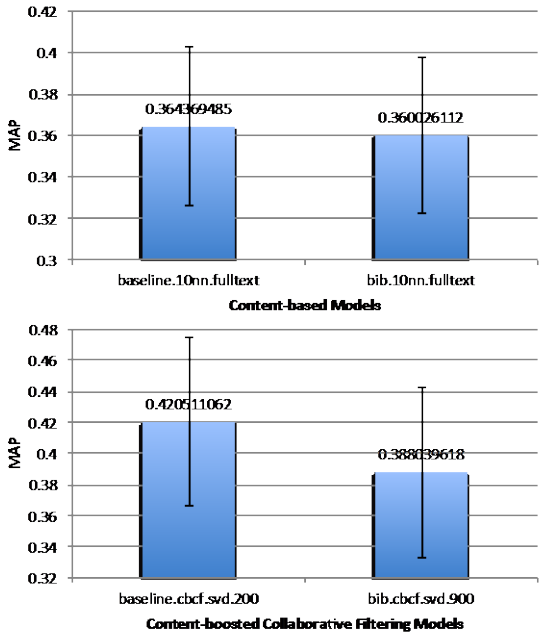


Figure 120: Bibliography-Augmented Models

The bibliography-augmented 10-NN.PO full-text model was chosen as the candidate for the bibliography-augmented CBF model because its mean MAP result was the highest among the other content-based models, as shown in the top left and bottom left of Figure **120**. The bibliography-augmented SVD content-boosted collaborative filtering model with 900 latent

topics, shown in the top right of Figure **120**, was selected as the representative of the bibliography-augmented CBCF model because its mean MAP result was the highest among the other CBCF models.

2) Test Set: Evaluation



Analysis of Variance (One-Way)

Summary

Groups	*Sample size*	*Sum*	*Mean*	*Variance*
baseline.10nn.fulltext	44	16.03226	0.36437	0.01679
bib.10nn.fulltext	44	15.84115	0.36003	0.01637

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*p-level*	*F crit*
Between Groups	0.00042	1	0.00042	0.02503	0.87466	5.61924
Within Groups	1.42586	86	0.01658			
Total	1.42628	87				

Analysis of Variance (One-Way)

Summary

Groups	*Sample size*	*Sum*	*Mean*	*Variance*
baseline.cbcf.svd.200	44	18.50249	0.42051	0.03358
bib.cbcf.svd.900	44	17.07374	0.38804	0.03472

ANOVA

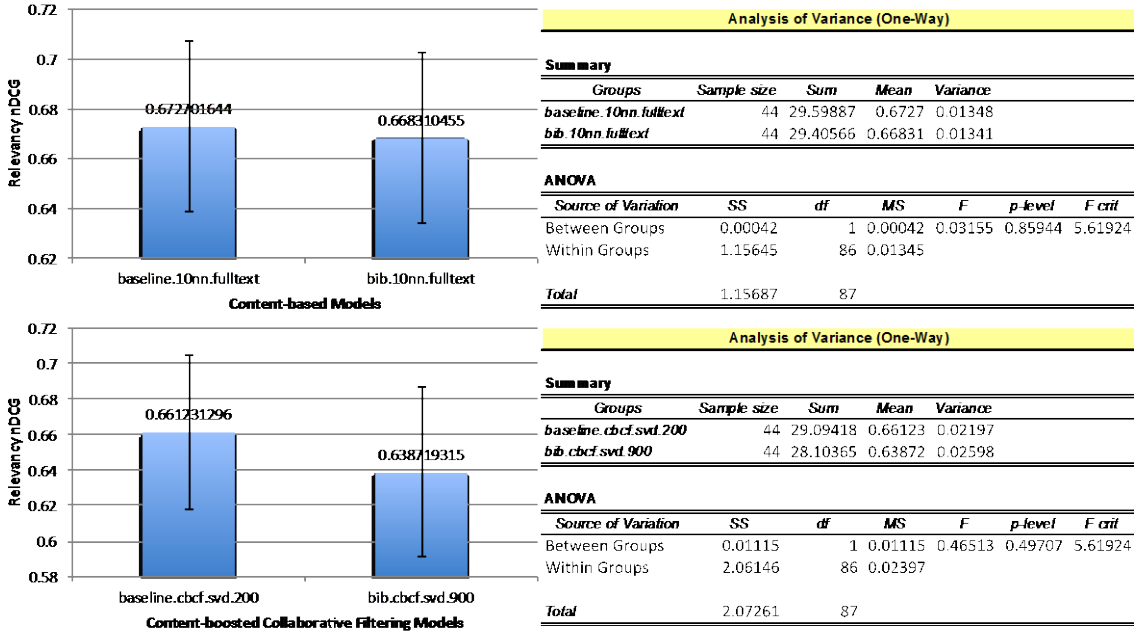
Source of Variation	*SS*	*df*	*MS*	*F*	*p-level*	*F crit*
Between Groups	0.0232	1	0.0232	0.67923	0.41213	5.61924
Within Groups	2.93704	86	0.03415			
Total	2.96024	87				

Figure 121: MAP Analysis of Bibliography Models in the Test Set

In the CBF comparison on MAP, both models were assessed with the 44 subjects who provided their publications in the test set. One-way ANOVA was applied to test the MAP results. At the top right of Figure **121**, the bibliography-augmented CBF model performed at approximately the same level as the CBF baseline.

In the CBCF comparison on MAP, the bibliography-augmented SVD content-boosted collaborative filtering model with 900 latent topics was assessed with the same 44 subjects. At

the bottom right of Figure **116**, the bibliography-augmented CBCF model performed lower than the CBCF baseline, but there was no statistically significant difference, using the p-value < 0.05 level of significance.



Analysis of Variance (One-Way)

Summary

Groups	*Sample size*	*Sum*	*Mean*	*Variance*
baseline.10nn.fulltext	44	29.59887	0.6727	0.01348
bib.10nn.fulltext	44	29.40566	0.66831	0.01341

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*p-level*	*F crit*
Between Groups	0.00042	1	0.00042	0.03155	0.85944	5.61924
Within Groups	1.15645	86	0.01345			
Total	1.15687	87				

Analysis of Variance (One-Way)

Summary

Groups	*Sample size*	*Sum*	*Mean*	*Variance*
baseline.cbcf.svd.200	44	29.09418	0.66123	0.02197
bib.cbcf.svd.900	44	28.10365	0.63872	0.02598

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*p-level*	*F crit*
Between Groups	0.01115	1	0.01115	0.46513	0.49707	5.61924
Within Groups	2.06146	86	0.02397			
Total	2.07261	87				

Figure 122: Relevance nDCG of Bibliography Models in the Test Set

In the CBF comparison of Relevance nDCG, both models were assessed with the 44 subjects who provided their publications in the test set. One-way ANOVA was applied to test the Relevance nDCG results. As shown in the top right of Figure **122**, the bibliography-augmented CBF model performed at approximately the same level as the CBF baseline.

In the CBCF comparison of Relevance nDCG, the bibliography-augmented SVD content-boosted collaborative filtering model with 900 latent topics was assessed with the same 44 subjects. As depicted in the bottom right of Figure **122**, the bibliography-augmented CBCF model

performed slightly lower than the CBCF baseline, but there was no statistically significant difference, using the p-value < 0.05 level of significance.

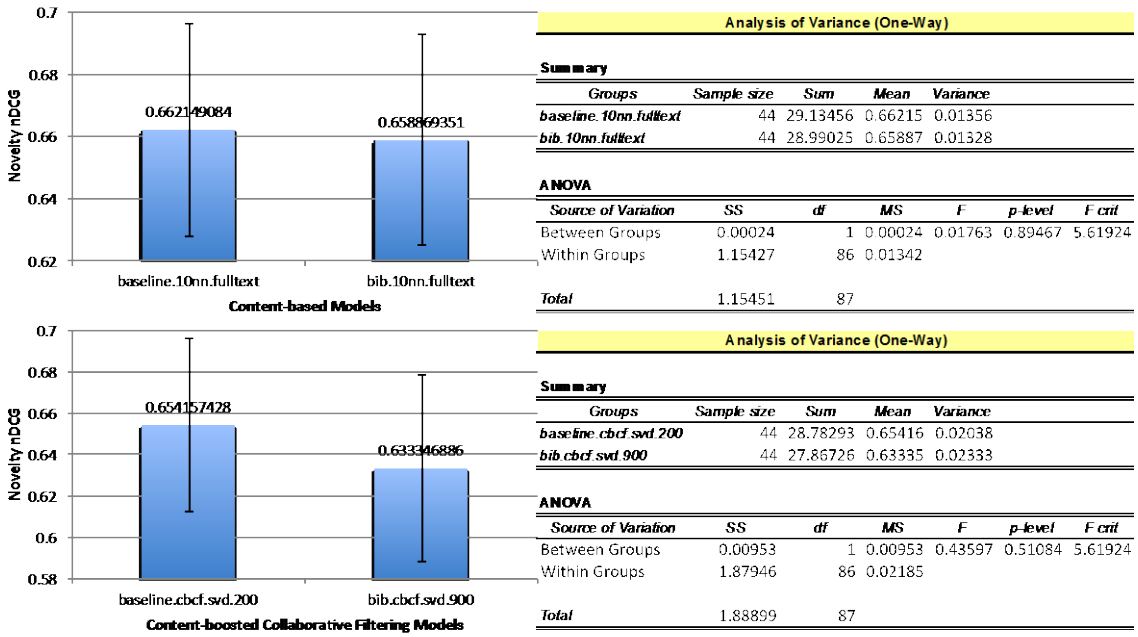


Figure 123: Novelty nDCG of Bibliography Models in the Test Set

In the CBF comparison of Novelty nDCG, both models were assessed with the 44 subjects who provided their publications in the test set. One-way ANOVA was applied to test the Relevance nDCG results. As shown in the top right of Figure **123**, the bibliography-augmented CBF model performed at approximately the same level as the CBF baseline.

In the CBCF comparison of Novelty nDCG, the bibliography-augmented SVD content-boosted collaborative filtering model with 900 latent topics was assessed with the same 44 subjects. As shown in the bottom right of Figure **123**, the bibliography-augmented CBCF model

performed marginally lower than the CBCF baseline, but there was no statistically significant difference, using the p-value < 0.05 level of significance.

10.5.1.3 Bookmarked Scholarly Papers (External Bookmark)

1) Training Set: Fitting Models

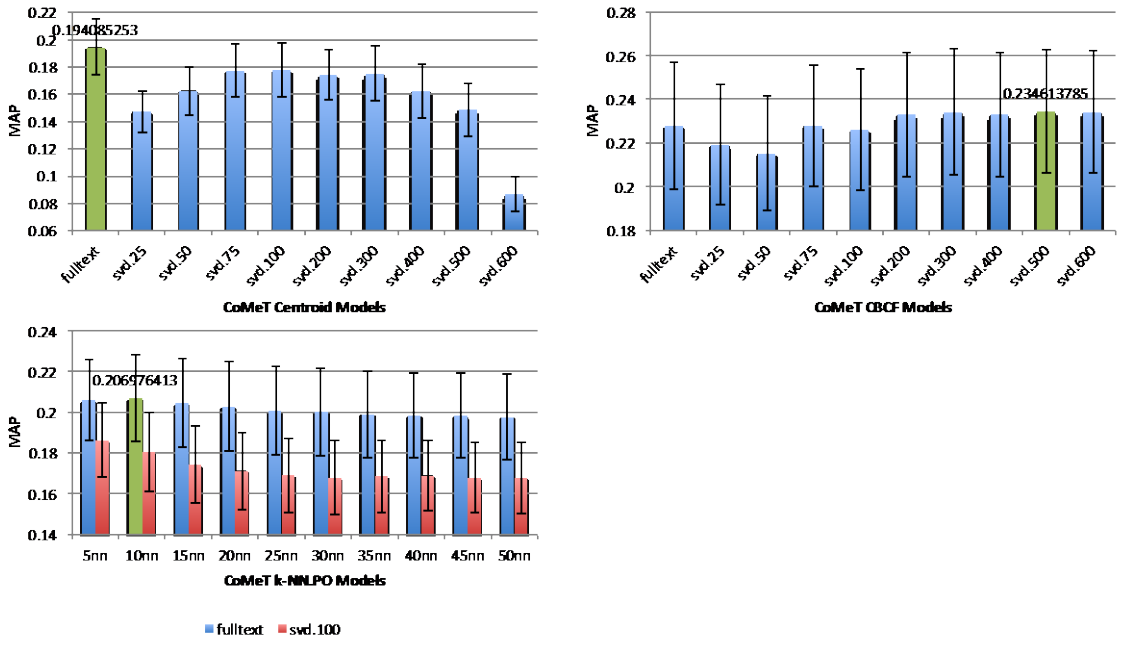


Figure 124: External-Bookmark-User Baseline Models

The CBF baselines were assessed in the training set with the 34 subjects who provided their external bookmark accounts or a list of bookmarked scientific articles, as shown in Figure **124**. The maximum MAP result of the individual centroid models was 0.19 on the full-text centroid model, which was the peak MAP result in comparison to the other SVD centroid models, as shown in the top left diagram in Figure **124**. The SVD centroid model with 100 latent topics

performed better than the other SVD models. Therefore, 100 latent topics were used later with the other SVD CBF baseline models. As shown in the bottom left of Figure **124**, all the KNN.PO baseline models performed steadily. The full-text KNN.PO models performed better than the SVD models. The maximum MAP result of the KNN.PO model was 0.21, in the 10-NN.PO full-text model.

The CBCF baseline models were assessed with the same 34 subjects in the training set, as depicted in the top right of Figure **124**. Their MAP results looked similarly, and the maximum MAP result of CBCF baseline models was in the CBCF SVD model with 200 latent topics.

The 10-NN.PO full-text baseline model was selected from the the others as a content-based baseline model, as shown in the top left and bottom left of Figure **124**. The SVD content-boosted collaborative filtering baseline with 200 latent topics was selected as the CBCF baseline representative, as shown in the top right of Figure **124**.

The external-bookmark-augmented 5-NN.PO full-text model was chosen as the representative of the external-bookmark-augmented CBF models because its mean MAP result was the highest among the other content-based models, as shown in the top left and bottom left of Figure **125**. The external-bookmark-augmented SVD content-boosted collaborative filtering model with 1500 latent topics, shown in the top right of Figure **125**, was selected as the representative of the external-bookmark-augmented CBCF models because its mean MAP result was the highest among the other CBCF models.

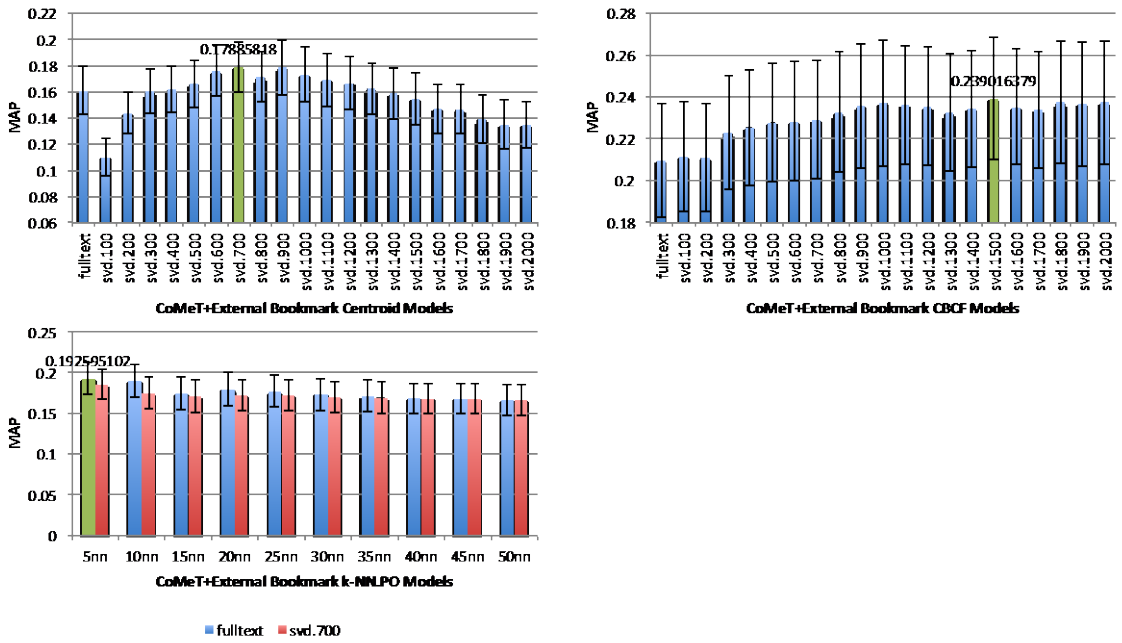
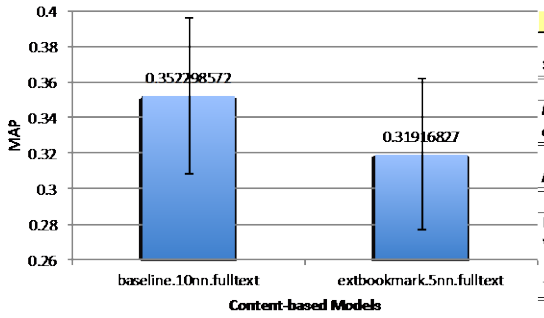


Figure 125: External-Bookmark-Augmented Models

2) Test Set: Evaluation

In the CBF comparison on MAP, both models were assessed with the 34 subjects in the test set. One-way ANOVA was applied to test the MAP results. From the top right of Figure **126**, the external-bookmark-augmented CBF model performed lower than the CBF baseline, but the difference was not statistically significance, using the p-value < 0.05 standard for significance.



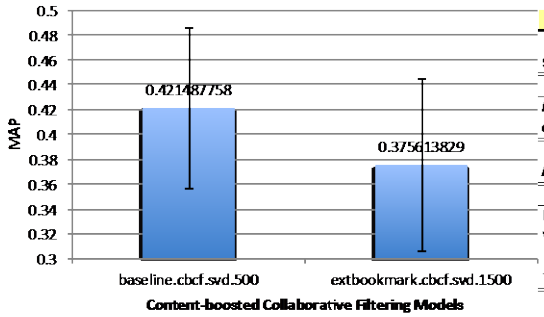
Analysis of Variance (One-Way)

Summary

Groups	Sample size	Sum	Mean	Variance
baseline.10nn.fulltext	34	11.97815	0.3523	0.01691
extbookmark.5nn.fulltext	34	10.85172	0.31917	0.01606

ANOVA

Source of Variation	SS	df	MS	F	p-level	F crit
Between Groups	0.01866	1	0.01866	1.13194	0.29124	5.68434
Within Groups	1.08798	66	0.01648			
Total	1.10664	67				



Analysis of Variance (One-Way)

Summary

Groups	Sample size	Sum	Mean	Variance
baseline.cbcf.svd.500	34	14.33058	0.42149	0.03731
extbookmark.cbcf.svd.1500	34	12.77087	0.37561	0.04239

ANOVA

Source of Variation	SS	df	MS	F	p-level	F crit
Between Groups	0.03578	1	0.03578	0.89778	0.34683	5.68434
Within Groups	2.63	66	0.03985			
Total	2.66577	67				

Figure 126: MAP Analysis of External Bookmark Models in the Test Set

In the CBCF comparison on MAP, the external-bookmark-augmented SVD content-boosted collaborative filtering model with 1500 latent topics was assessed with the same 34 subjects. From the bottom right of Figure **126**, the external-bookmark-augmented CBCF model performed lower than the CBCF baseline, but the difference was not statistically significant, using the p-value < 0.05 level of significance.

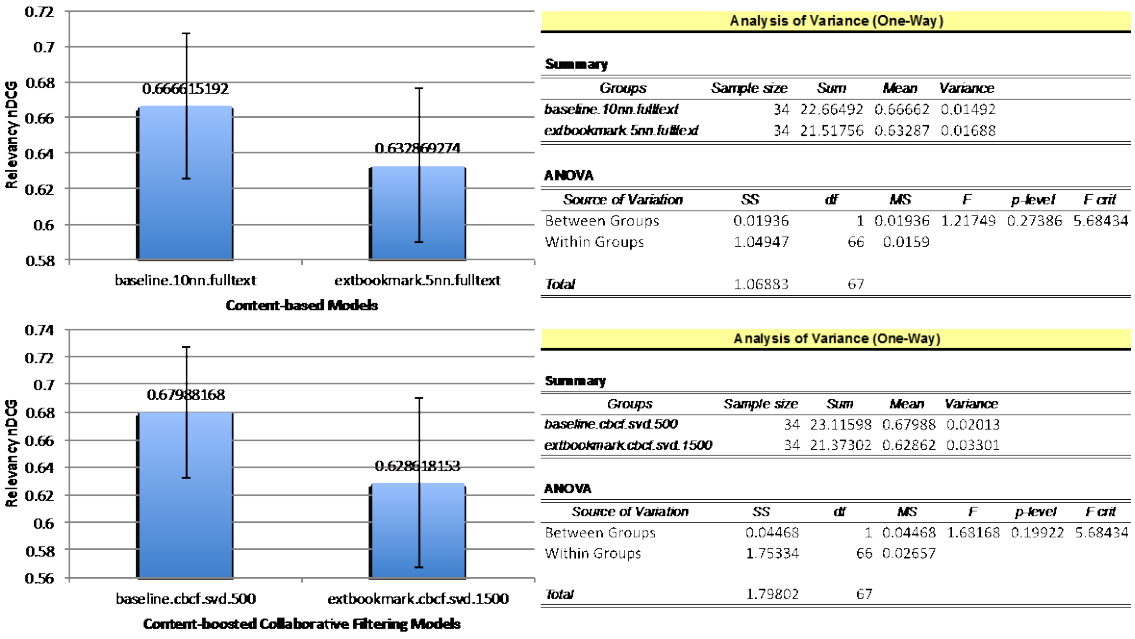


Figure 127: Relevance nDCG of External Bookmark Models in the Test Set

In the CBF comparison on Relevance nDCG, both models were assessed with the 34 subjects in the test set. One-way ANOVA was applied to test the Relevance nDCG results. As shown in the top right of Figure **127**, the external-bookmark-augmented CBF model performed lower than with the CBF baseline, but there was no significant difference, using the p-value < 0.05 level of significance.

In the CBCF comparison of Relevance nDCG, the external-bookmark-augmented SVD content-boosted collaborative filtering model with 1500 latent topics was assessed with the same 34 subjects. As shown in the bottom right of Figure **127**, the external-bookmark-augmented CBCF model performed lower than the CBCF baseline, but the difference was not statistically significant, using the p-value < 0.05 standard for significance.

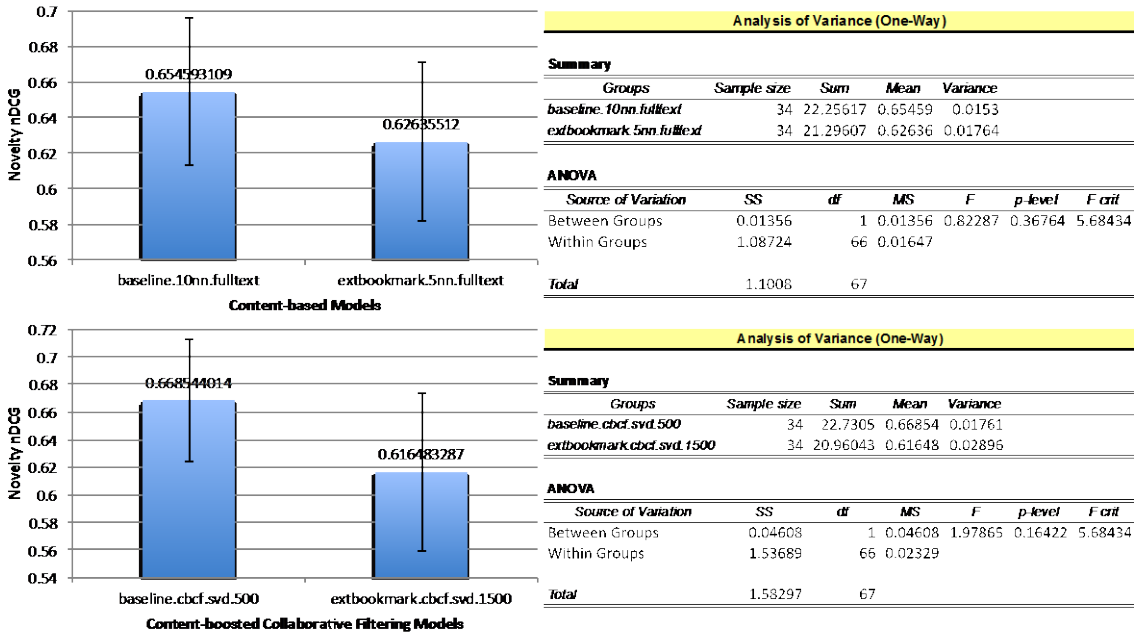


Figure 128: Novelty nDCG of External Bookmark Models in the Test Set

In the CBF comparison of Novelty nDCG, both models were assessed with the 34 subjects in the test set. One-way ANOVA was applied to test the Novelty nDCG results. As shown in the top right of Figure **128**, the external-bookmark-augmented CBF model performed lower than with the CBF baseline, but there was no significant difference, using the p-value < 0.05 standard for significance.

In the CBCF comparison of Novelty nDCG, the external-bookmark-augmented SVD content-boosted collaborative filtering model with 1500 latent topics was assessed with the same 34 subjects. As shown in the bottom right of Figure **128**, the external-bookmark-augmented CBCF model performed lower than the CBCF baseline, but there was no significant difference, using the p-value < 0.05 standard for significance.

10.5.2 Study 4.2: Cold-Start Context

The six external-source-augmented experimental models and the six baseline models from the previous section were carried away to assess on the cold-start problem study. In order to study the effect of cold-start problems, 20 bookmarked talks were randomly selected from the training set for each subject. These talks were divided into 21 different-sized bins, ranging from, no bookmarks at all, to 20 bookmarks in the training set. The numbers of bins to which users were assigned was based on the bookmarks they had in their user profiles. The random selection was repeated 10 “round” times. The evaluation result was the average result of these 10 “round” iterations.

10.5.2.1 Homepage

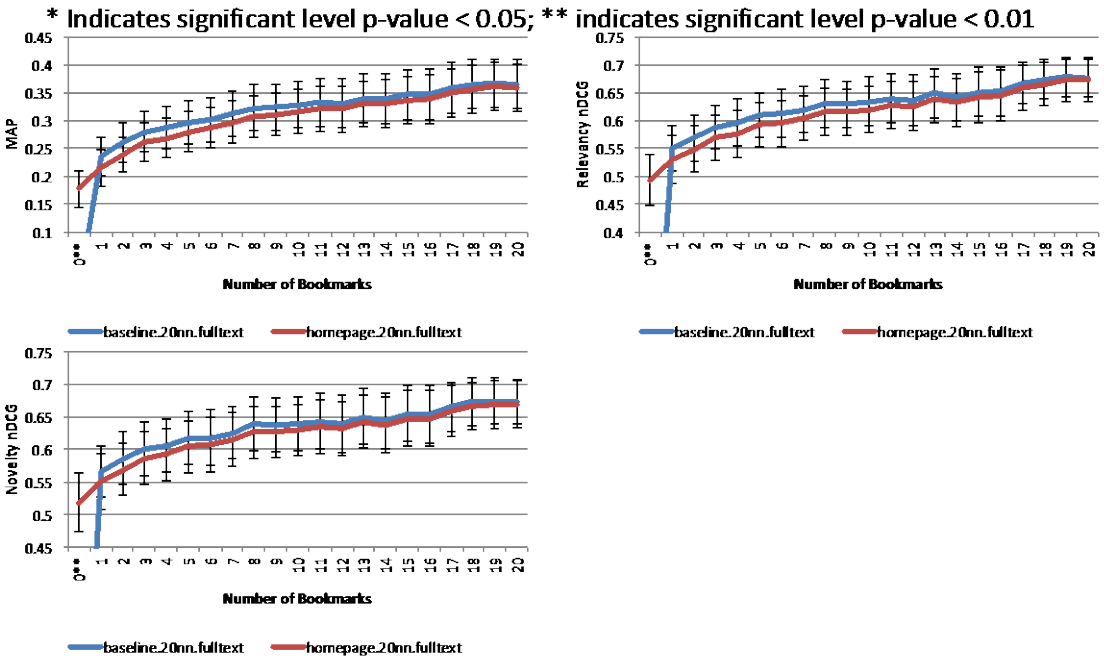


Figure 129: Cold-Start Homepage CBF Models in the Test Set

In this cold-start problem, the homepage-augmented 20-NN.PO full-text model and the homepage-augmented SVD CBCF model with 500 latent topics were evaluated against the 20-NN.PO full-text baseline one and the SVD baseline CBCF one with 200 latent topics, respectively.

These four models were assessed with 29 subjects who provided their homepage information. One-way ANOVA was applied to test the MAP, Relevance nDCG, and Novelty nDCG results. The results showed that the homepage augmentation recommendations on CBF models improved significantly better than the baselines in the initial cold-start situation, in which users had no bookmarks in any of the three measures. After that, the homepage-augmented CBF

model performed slightly lower than the baselines, but there was no statistically significant difference in any of the three measures.

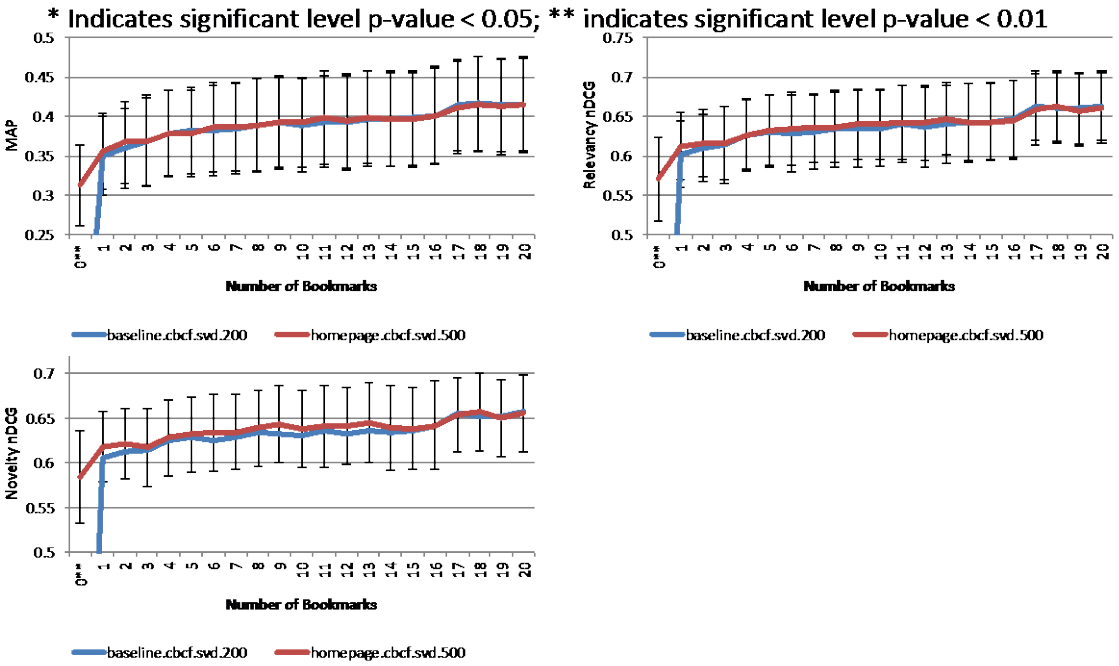


Figure 130: Cold-Start Homepage CBCF Models in the Test Set

The homepage-augmented CBCF model performed at the same level as the baseline did in the most stages of the cold-start situations, except when users had no bookmarks.

10.5.2.2 Bibliography

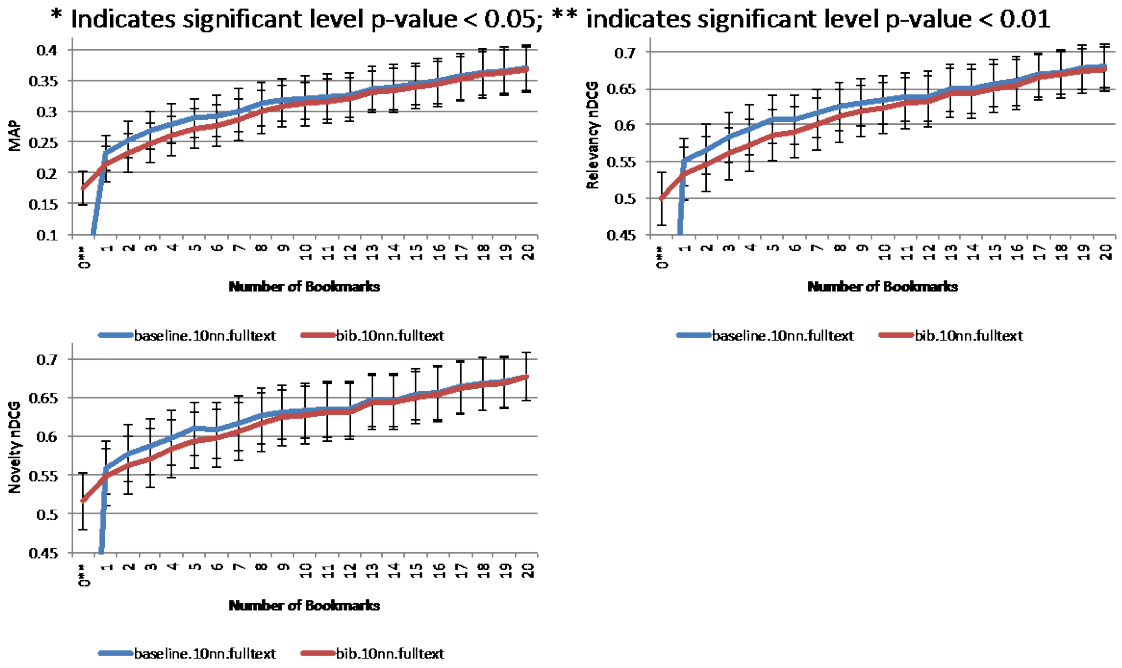


Figure 131: Bibliography CBF Models for the Cold-Start Problem in the Test Set

In this cold-start problem, the bibliography-augmented 10-NN.PO full-text model was evaluated against the 10-NN.PO full-text baseline model, and the bibliography-augmented SVD CBCF model with 900 latent topics was evaluated against the SVD baseline CBCF model with 200 latent topics.

These four models were assessed with the 44 subjects who provided their publications. One-way ANOVA was applied to test the MAP, Relevance nDCG, and Novelty nDCG results. The results showed that the bibliography augmentation recommendations on the CBF models improved significantly over the baselines in the initial cold-start situation, in which users had no bookmarks in any of the three measures. After that, the bibliography-augmented CBF one

performed slightly lower than the baselines but there was no statistically significant difference in any of the three measures.

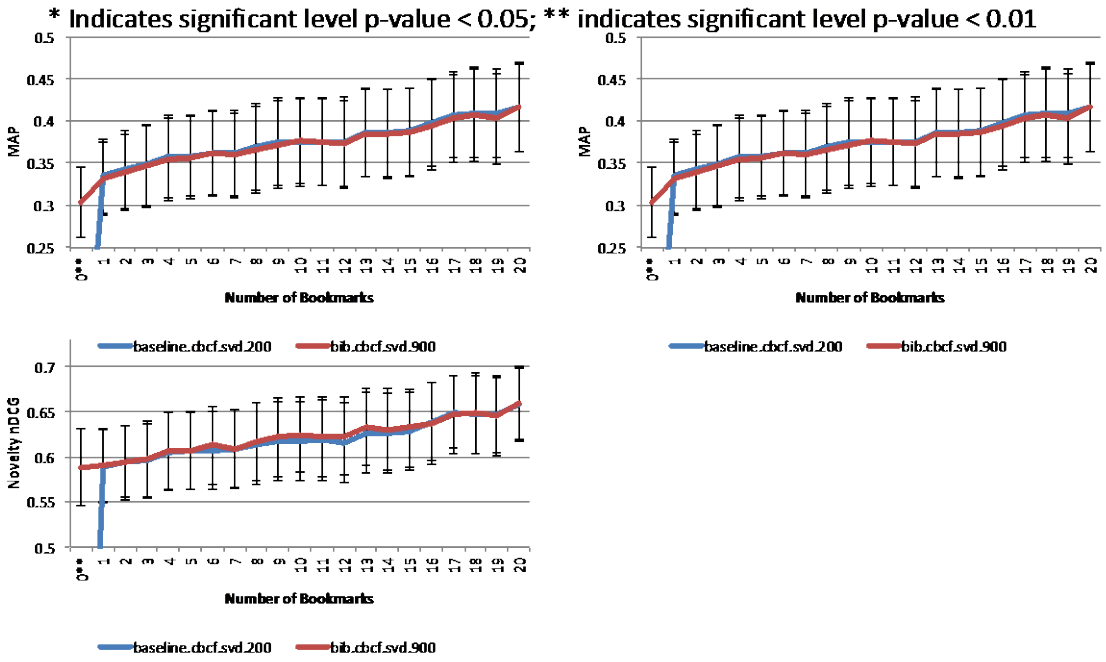


Figure 132: Cold-Start Bibliography CBCF Models in the Test Set

The bibliography-augmented CBCF model performed at the same level as the baseline did in the most stages of the cold-start situations except in the first stage when users have no bookmark.

10.5.2.3 External Bookmark

In this cold-start problem, the external-bookmark-augmented 5-NN.PO full-text model was evaluated against the 10-NN.PO full-text baseline model, and the external-bookmark-augmented

SVD CBCF model with 1,500 latent topics was evaluated against the SVD baseline CBCF one with 500 latent topics.

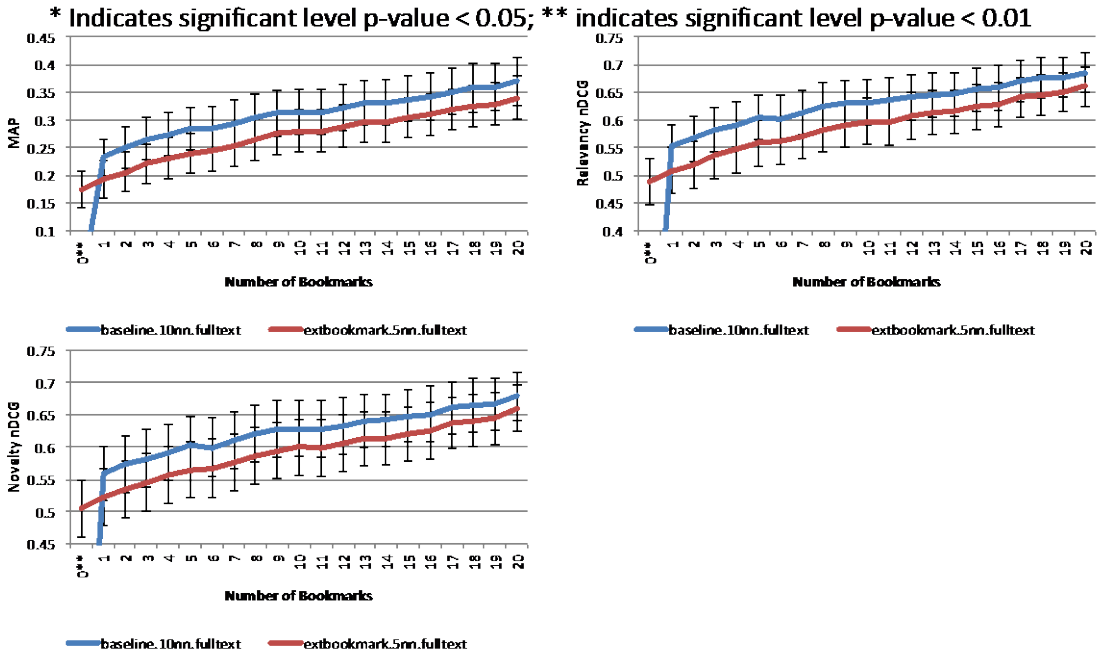


Figure 133: External Bookmark CBF Models for the Cold-Start Problem in the Test Set

These four models were assessed with the 34 subjects who provided their external bookmark accounts or a list of bookmarked scientific articles. One-way ANOVA was applied to test the MAP, Relevance nDCG, and Novelty nDCG results. The results showed that the external bookmark augmentation recommendations on the CBF models improved significantly over the baselines in the initial cold-start situation, in which users had no bookmarks in any of the three measures. After that, the external-bookmark-augmented CBF model performed slightly lower than the baselines, but there was no statistically significant difference in any of the three measures.

The external-bookmark-augmented CBCF model performed at the same level as the baseline in the most stages of cold-start situations, except in the first stage, when users had no bookmarks.

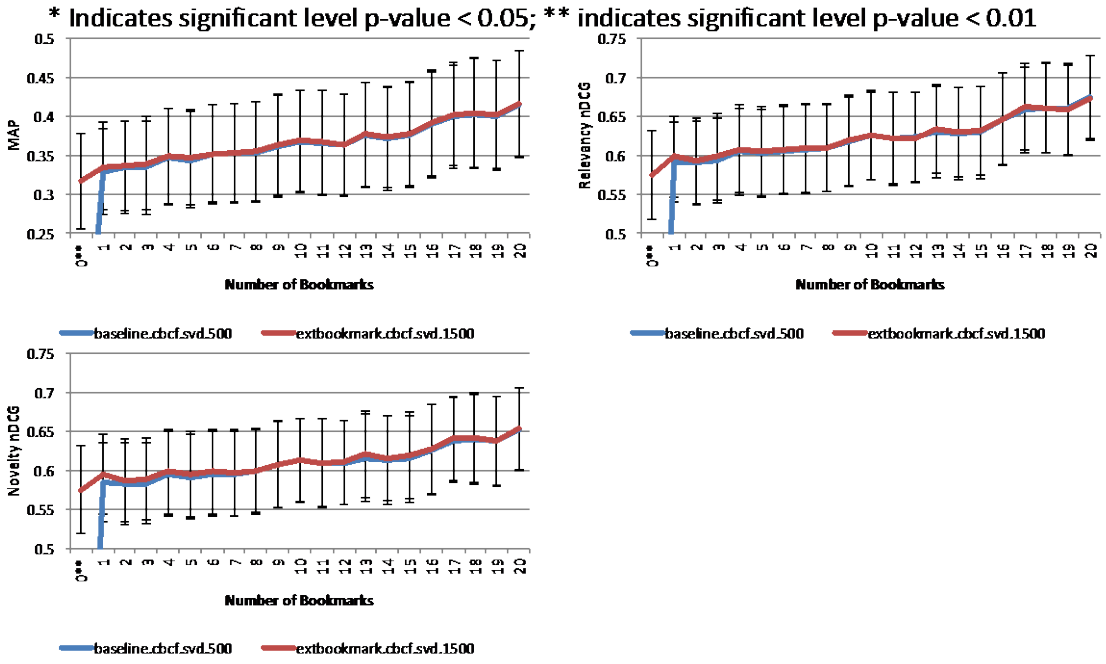


Figure 134: Cold-Start External Bookmark CBCF Models in the Test Set

10.5.3 Study 4.3: Recommendation Fusion

The six external-source-augmented experimental models and the six baseline models from study 4.1 were carried away to assess in this experiment. The bookmarked CoMeT talks in the training set were used to construct the user models. The bookmarked talks in the test set were used to evaluate the performance of the models.

10.5.3.1 Different-Source-Different-Approach Fusion

1) Homepage and Bibliography Fusion

The experimental CombMNZ and CombSUM approaches fusing of the homepage-augmented representative model and the bibliography-augmented representative model were assessed with the 29 subjects who provided both their homepages and their publications. These models were evaluated against the CBF baseline (full-text 20-NN.PO model) and CBCF baseline (SVD CBCF model with 200 latent topics).

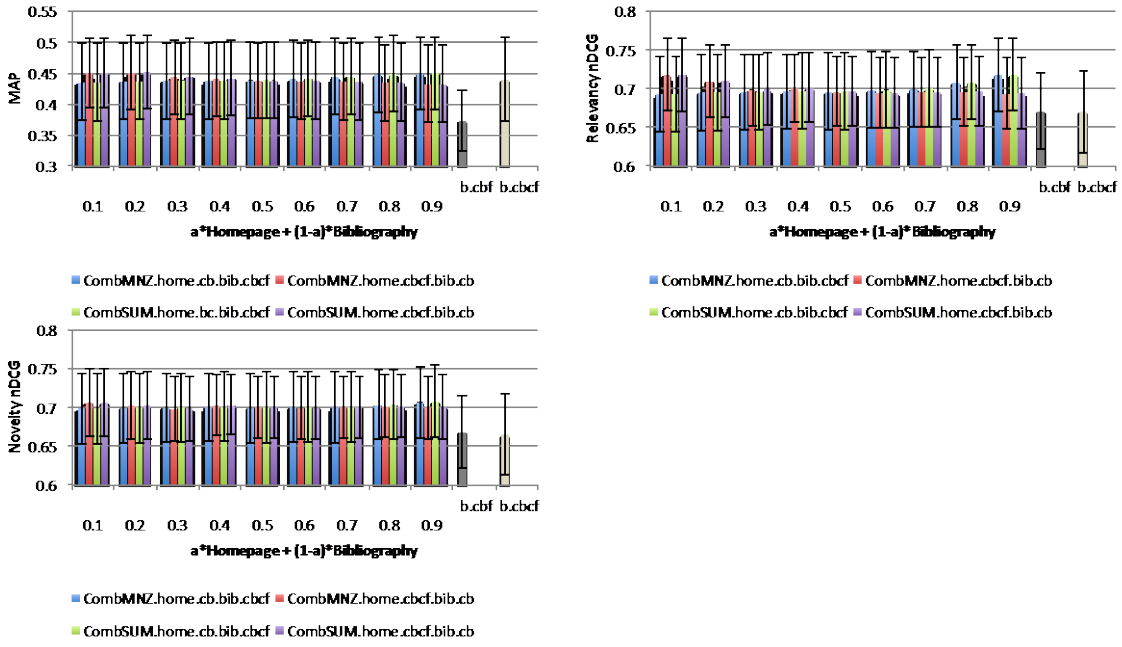


Figure 135: Homepage + Bibliography Fusion

One-way ANOVA with Tuky's HSD Post Hoc test was applied to test the MAP, Relevance nDCG, and Novelty nDCG results. The results of the recommendation fusion of the homepage-augmented model and the bibliography-augmented model varied from 0.43 to 0.45 for

MAP, from 0.69 to 0.72 for Relevance nDCG, and from 0.70 to 0.71 for Novelty nDCG, depending on the type of fusion, the recommending method, and the weight of fusion. Models fusing with both the CombMNZ method and the CombSUM method performed well in any stepping weight.

In the MAP measurements, all 36 homepage-bibliography-fusion models (18 CombMNZ models and 18 CombSUM models) performed slightly better than the CBF baseline, and at the same level as CBCF baseline. In the Relevance and Novelty nDCG measures, all 36 homepage-bibliography-fusion models (18 CombMNZ models and 18 CombSUM models) performed slightly better than both the CBF and the CBCF baselines, but the differences were not significant.

2) Homepage and External Bookmark Fusion

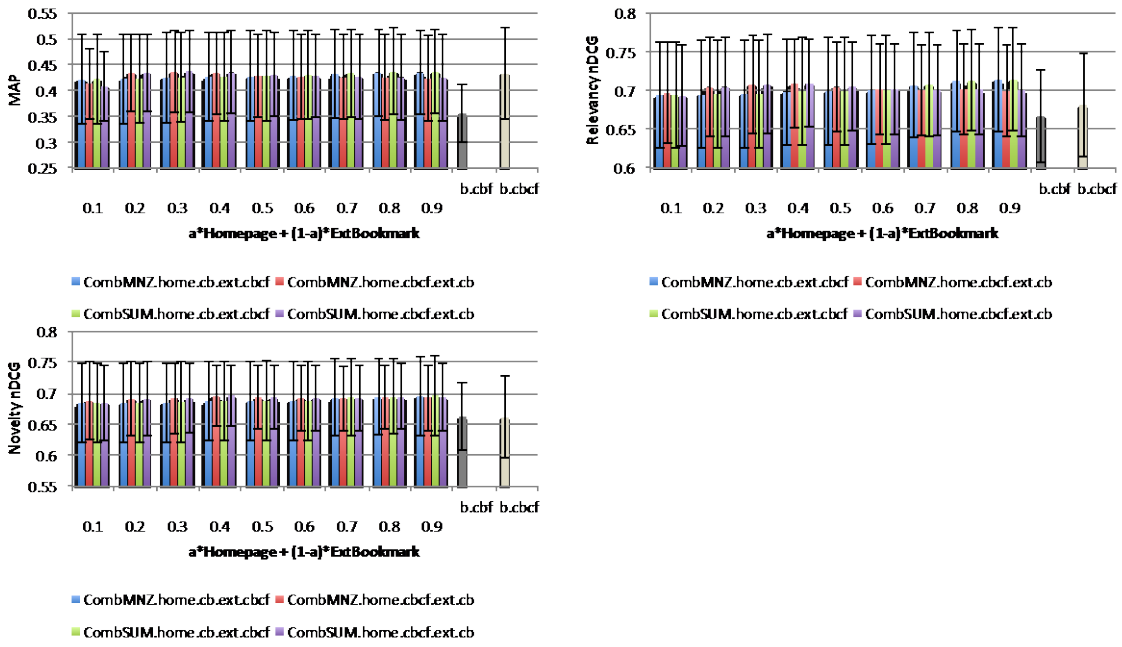


Figure 136: Homepage + External Bookmark Fusion

The experimental CombMNZ and CombSUM approaches fusing of the homepage-augmented representative model and the external-bookmark-augmented representative model were assessed with the 20 subjects who provided both their homepages, and their external bookmark accounts or a list of bookmarked scientific articles. These models were evaluated against the CBF baseline (full-text 10-NN.PO model) and the CBCF baseline (SVD CBCF model with 200 latent-topics).

One-way ANOVA with Tuky's HSD Post Hoc test was applied to test the MAP, Relevance nDCG, and Novelty nDCG results. The results of the recommendation fusion of the homepage-augmented model and the external-bookmark-augmented model varied from 0.40 to 0.42 for MAP, from 0.69 to 0.71 for Relevance nDCG, and from 0.68 to 0.70 for Novelty nDCG, depending on the type of fusion, the recommending method, and the weight of fusion. Models fusing with both the CombMNZ method and CombSUM method performed well in any stepping weight.

In the MAP measurements, all 36 homepage-external-bookmark-fusion models (18 CombMNZ models and 18 CombSUM models) performed slightly better than the CBF baseline, and at the same level as the CBCF baseline. In the Relevance and Novelty nDCG measures, all 36 homepage-external-bookmark-fusion models (18 CombMNZ models and 18 CombSUM models) performed slightly better than both the CBF and the CBCF baselines, but the difference was not statistically significant.

3) Bibliography and External Bookmark Fusion

The experimental CombMNZ and CombSUM approaches fusing of the bibliography-augmented representative model and the external-bookmark-augmented representative model were assessed with the 34 subjects who provided both their publications, and their external

bookmark accounts or a list of bookmarked scientific articles. These models were evaluated against the CBF baseline (full-text 10-NN.PO model), and the CBCF baseline (SVD CBCF model with 500 latent topics).

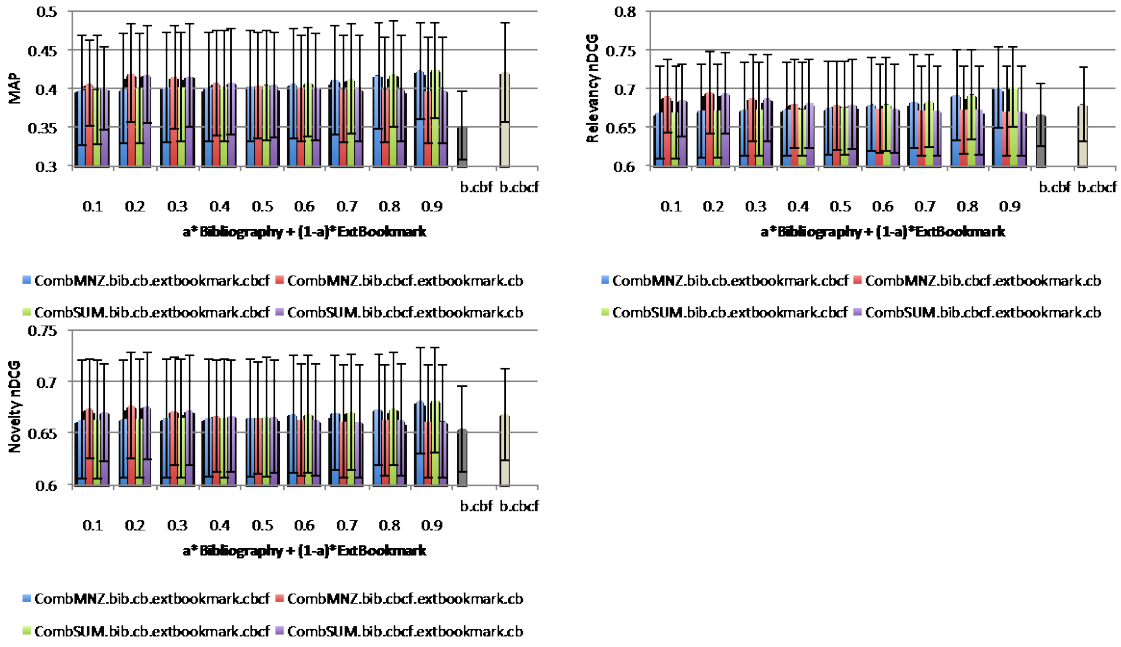


Figure 137: Bibliography + External Bookmark Fusion

One-way ANOVA with Tuky's HSD Post Hoc test was applied to test the MAP, Relevance nDCG, and Novelty nDCG results. The results of the recommendation fusion of the homepage-augmented model and the external-bookmark-augmented model varied from 0.40 to 0.42 for MAP, from 0.67 to 0.70 for Relevance nDCG, and from 0.66 to 0.68 for Novelty nDCG, depending on the type of fusion, the recommending method, and the weight of fusion. Models fusing with both the CombMNZ method and the CombSUM method performed well in any stepping weight.

In the MAP measurements, all 36 bibliography-external-bookmark-fusion models (18 CombMNZ models and 18 CombSUM models) performed slightly better than the CBF baseline, and at the same level as the CBCF baseline. In Relevance and Novelty nDCG measures, all 36 bibliography-external-bookmark-fusion models (18 CombMNZ models and 18 CombSUM models) performed slightly better than both the CBF and the CBCF baselines, but the differences were not significant.

10.5.3.2 Same-Source Fusion

1) Homepage

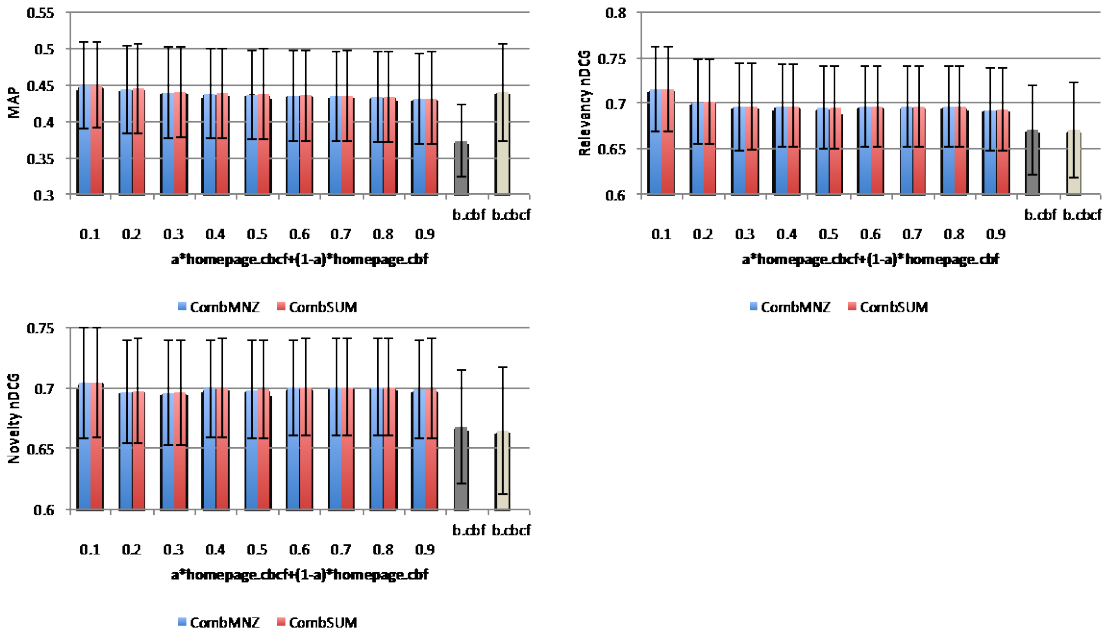


Figure 138: Homepage Same-Source Fusion Results

The recommendations from the two homepage augmentation models, the homepage-augmented full-text 20-NN.PO model and the homepage-augmented SVD CBCF model with 500 latent topics, were combined and evaluated against two baselines, which were the full-text 20-NN.PO baseline CBF model and the SVD baseline CBCF model with 200 latent topics. These four models were assessed with the 29 subjects who provided their homepage information.

One-way ANOVA with Tuky's HSD Post Hoc test was applied to test the MAP, Relevance nDCG, and Novelty nDCG results. The results of the experimental same-source-different-method homepage-augmented recommendation fusion models varied from 0.43 to 0.45 for MAP, from 0.69 to 0.71 for Relevance nDCG, and from 0.69 to 0.70 for Novelty nDCG, depending on the type of fusion, the recommending method, and the weight of fusion. Models fusing with both methods performed well in any stepping weights.

In the MAP measurements, all 18 same-source-different-method homepage-augmented models (nine CombMNZ models and nine CombSUM models) slightly better than the CBF baseline, and at the same level as the CBCF baseline. In Relevance and Novelty nDCG measures, all 18 bibliography-external-bookmark-fusion models (18 CombMNZ models and 18 CombSUM models) performed slightly better than both the CBF and the CBCF baselines, but the differnces in their performance were not significant.

2) Bibliography

The recommendations from the two bibliography augmentation models, the bibliography-augmented full-text 10-NN.PO model and the bibliography-augmented SVD CBCF model with 900 latent topics, were combined and evaluated against two baselines, which were the full-text

10-NN.PO baseline CBF model and the SVD baseline CBCF model with 200 latent topics. These four models were assessed with the 44 subjects who provided their publications.

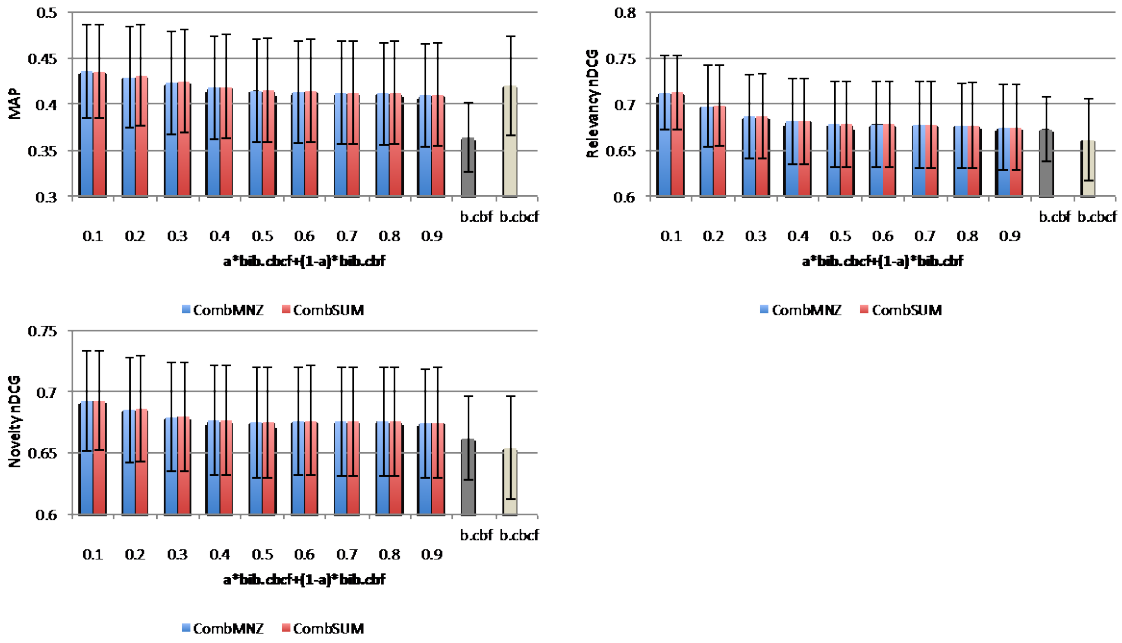


Figure 139: Bibliography Same-Source Fusion Results

One-way ANOVA with Tuky's HSD Post Hoc test was applied to test the MAP, Relevance nDCG, and Novelty nDCG results. The results of the experimental same-source-different-method bibliography-augmented recommendation fusion models varied from 0.41 to 0.44 for MAP, from 0.67 to 0.71 for Relevance nDCG, and from 0.67 to 0.69 for Novelty nDCG, depending on the type of fusion, the recommending method, and the weight of fusion. Models fusing with both methods performed well in any stepping weights.

In the MAP measurements, all 18 same-source-different-method homepage-augmented models (nine CombMNZ models and nine CombSUM models) performed slightly better than the

CBF baseline, and at the same level as the CBCF baseline. In Relevance and Novelty nDCG measures, all 18 bibliography-external-bookmark-fusion models (nine CombMNZ models and nine CombSUM models) performed slightly better than both the CBF and the CBCF baselines, but there were no significant differences in their performance.

3) External Bookmark

The recommendations from the two external bookmark augmentation models, the external-bookmark-augmented full-text 5-NN.PO model and the bibliography-augmented SVD CBCF model with 1,500 latent topics, were combined and evaluated against two baselines, which were the full-text 10-NN.PO baseline CBF model and the SVD baseline CBCF model with 500 latent topics. These four models were assessed with 34 subjects who provided their external bookmark accounts or a list of bookmarked scientific articles.

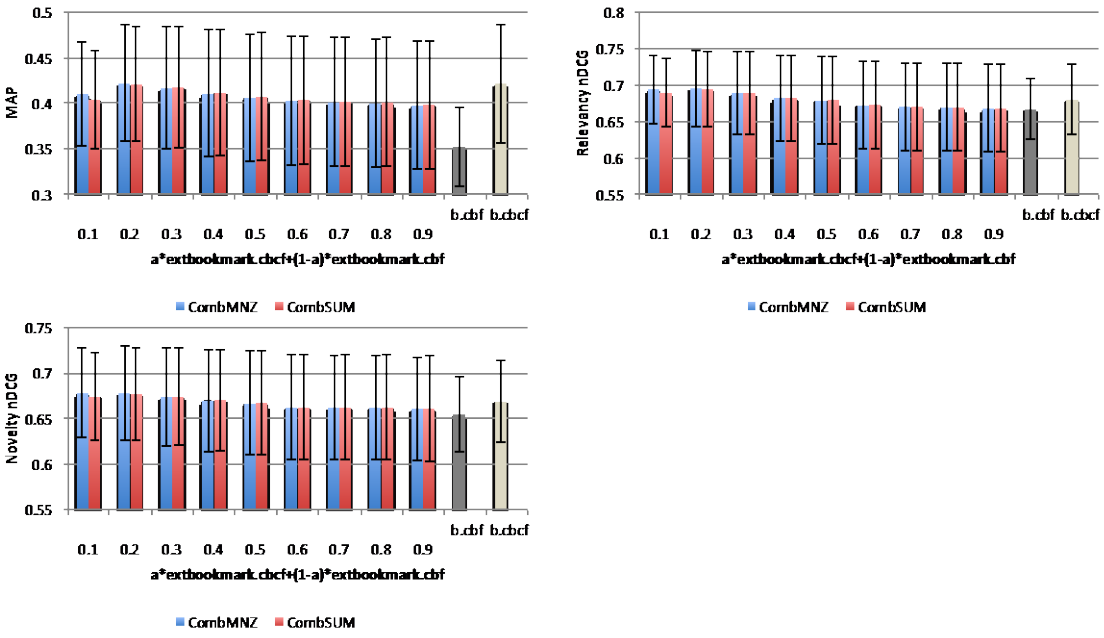


Figure 140: External-Bookmark Same-Source Fusion Results

One-way ANOVA with Tuky's HSD Post Hoc test was applied to test the MAP, Relevance nDCG, and Novelty nDCG results. The results of the experimental same-source-different-method bibliography-augmented recommendation fusion models varied from 0.40 to 0.42 for MAP, from 0.67 to 0.69 for Relevance nDCG, and from 0.66 to 0.68 for Novelty nDCG, depending on the type of fusion, the recommending method, and the weight of fusion. Models fusing with both methods performed well in any stepping weights.

In the MAP measurements, all 18 same-source-different-method external-bookmark-augmented models (nine CombMNZ models and nine CombSUM models) performed slightly better than the CBF baseline, and at the same level as the CBCF baseline. In Relevance and Novelty nDCG measures, all 18 bibliography-external-bookmark-fusion models (nine CombMNZ models and nine CombSUM models) performed at approximately the same level as both the CBF and the CBCF baselines.

10.6 SUMMARY AND DISCUSSION

In this chapter, a user study was conducted to collect the CoMeT dataset. With this dataset, six external-source-augmented models were selected from the external-source-augmented models used in the experiment run in the training set. These models were assessed with 44 participants. The validation processes were repeated in the test set for all three sub-studies, which were: Study 4.1, external source augmentation improvement; Study 4.2, the cold-start problem; and Study 4.3, the recommendation fusion.

In Study 4.1, which concerned external source augmentation improvement, none of the augmented models improved on the performance of the baselines. Unlike the CN3 studies, augmenting the external information into the CoMeT user model does not provide any recommendation benefit in any measure.

In the study 4.2, which focused on the cold-start problem, experimental models augmented with either a homepage, a bibliography, or an external-bookmark external source helped improve recommendation results significantly in all three measures, but only when users had no bookmarked talks, as was the case in the cold-start problem study in Chapter 7.0 . However, after the user profiles started to have bookmarked talks, the baseline models performed slightly better than the augmented CBF and CBCF models in all three measures.

In Study 4.3, which concerned recommendation fusion, experimental models fusing with either a CombMNZ or a CombSUM approach performed at roughly the same MAP level as the CBCF baselines. They also slightly improved their MAP performance compared to the CBF baselines, but the difference was not significant. In the Relevance and Novelty nDCG measures, the homepage-bibliography-fusion and homepage-external-bookmark-fusion models slightly improved their performance over the CBF and CBCF baselines. However, the bibliography-external-bookmark-fusion models did not improve over the baselines in either the Relevance or the Novelty nDCG measures.

Discussion

The results of CoMeT studies did not perform differently from those of the CN3 studies. It is speculated that this is because of the difference in data collection between two systems, and the two different assumptions of their ground truth. The CN3 dataset was sparse. Most of CN3 users

bookmarked talks from conferences they attended. It was assumed that the remaining talks left in the corpus did not interest the users. The users had no real chance to view and judge these talks. On the contrary, in the CoMeT dataset, participants had to provide feedback on every single talk in both the training and the test set by deciding whether or not to bookmark. Thus, the CoMeT dataset had a complete ground truth. Based on the complete ground truth, we hypothesized that there was a sufficient amount of bookmarked talks in the CoMeT study to construct good baseline models. Consequently, the MAP results of the baseline models in all three CoMeT sub-studies were already high. Adding extra information into user profiles did not help to improve the performance of the models. Secondly, we hypothesized that this was because of the different natures of the two systems. CN3 hosts talks given at research conferences. The CN3 talks tend to be more focused and interrelated, while the CoMeT talks were more diversified. The interest areas of the CoMeT talks range from computer sciences, to medical and health sciences, to liberal arts and philosophy. Another observation is that the bibliography source helped improve the models' performance in the CN3 studies, but it did not do so in the CoMeT studies because the differences in the nature of the participants. Most of the CN3 users were scholars or senior researchers. They had already published a considerable number of papers (median: 27). In contrast, the CoMeT users were PhD students or junior researchers, and they had published a smaller number of papers (median: 9). Coupled with the nature of the CN3 system, this meant that the CN3 users were more likely to be more experienced, and tended to choose to attend on talks related to their research interests. On the other hand, the CoMeT users tended to attend talks that explored topics outside their research interests.

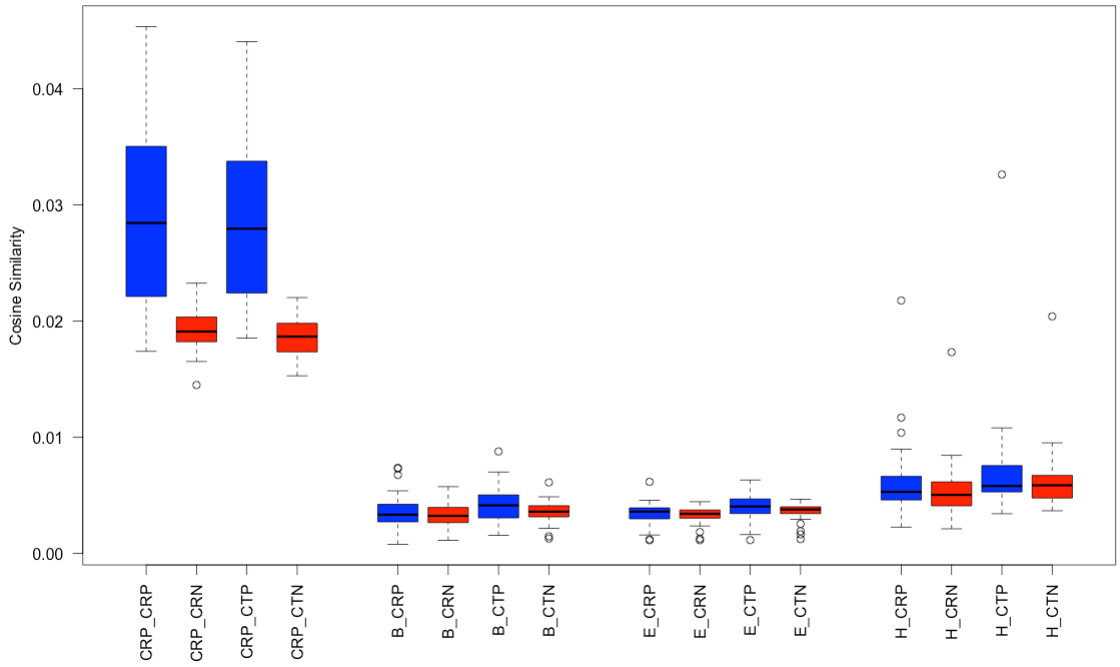


Figure 141: Full-Text Similarities between Bookmarked CoMeT Talks and External Sources

The results of the content-based recommendation with external source augmentation (CBF) study showed that there was no experimental model that significantly outperformed over the CBF baselines. The first four boxplots from the left inward in Figure **141** and Figure **142** examine in greater detail of the similarity between each individual CoMeT talk and the other three external-source documents. As shown in the first four boxplots from the left, bookmarked CoMeT talks are closer to one another in both the training set (the first blue boxplot from the left) and the test set (the second blue boxplot from the left), and further away from non-bookmarked talks in both the training set (the first red boxplot from the left) and the test set (the second red boxplot from the left), in both the full-text and the SVD models.

Bibliography and external-bookmark documents were not a good indicator to differentiate the bookmarked talks from the non-bookmarked ones, as shown in the second set of four boxplots and the third set from the left in both Figure **141** and Figure **142**. However, homepage documents (the fourth set of four boxplots from the left) seemed to have a few more different similarities between the bookmarked talks and the non-bookmarked ones. The gaps between them were not big enough when compared to CoMeT documents only. We hypothesized that the homepage documents attracted greater general interest, but they were also contaminated with many irrelevant terms.

Another hypothesis was that the external-source documents were in a different space of the CoMeT domain. When attending academic talks, users would go to the talks if they had spare time, and they were more likely to join the talks that shared their general interests. As a result, the bibliography-augmented CBF approach, which significantly outperformed the baseline in the CN3 study, did not work well in this user study.

Another hypothesis was that in a conference attending situation, most of the CN3 users were senior researchers. They had a considerable number of publications, whereas most of the CoMeT users were PhD students or junior researchers, and most of them had fewer than 20 publications. As a result, the bibliography-augmented CBF models did not perform in the same way as the ones in the CN3 study.

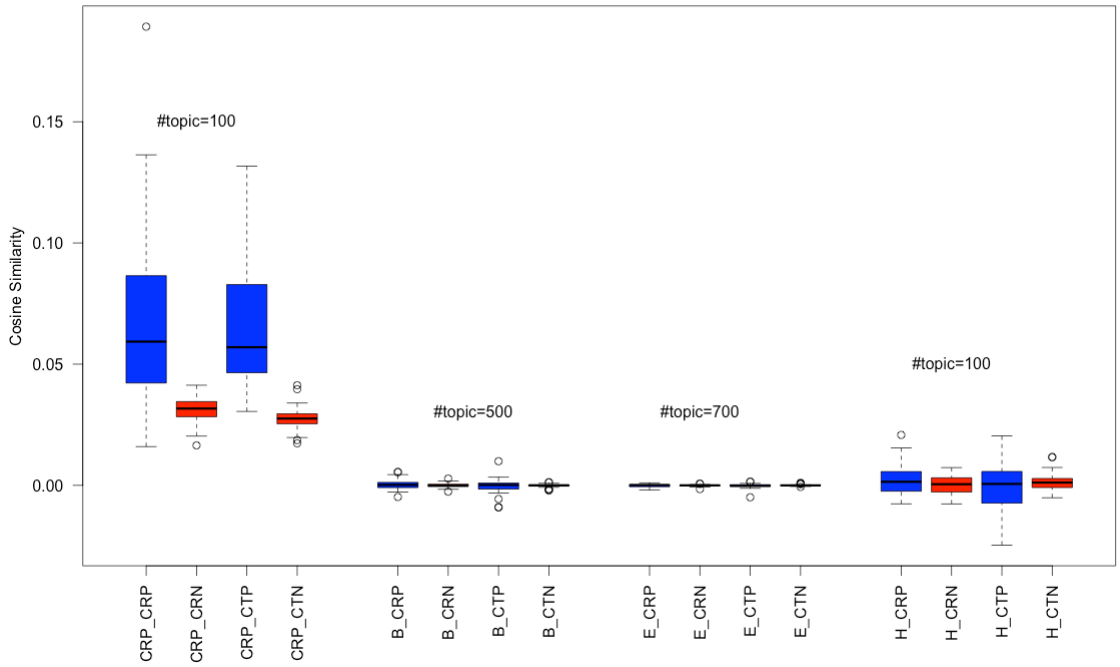


Figure 142: Latent Semantic Similarities between Bookmarked CoMeT Talks and External Sources

For CBCF study, the results were analyzed by examining the similarity between the user profile of the baseline models and the user profile of the experimental models with external source augmentation, as shown in Figure **143**. The CBCF baselines performed very well due to a high similarity between users and their 10 nearest peers in both the training set (the first blue boxplot from the left) and the test set (the first red boxplot form the left). The two particular external sources types, the bibliography (the second blue and red boxplots form the left) and the external bookmarked scholarly papers (the third blue and red boxplots form the left), did not have any impact on the user similarity matrix, nor did they improve the recommendation results in the global-impact situation. In fact, they reduced the similarity between users and their peers, except in the homepage-augmented models. As expected, in Study 4.1, the SVD CBCF baselines

performed slightly higher in all three measures than the unigram vector space CBCF models with any external source augmentation.

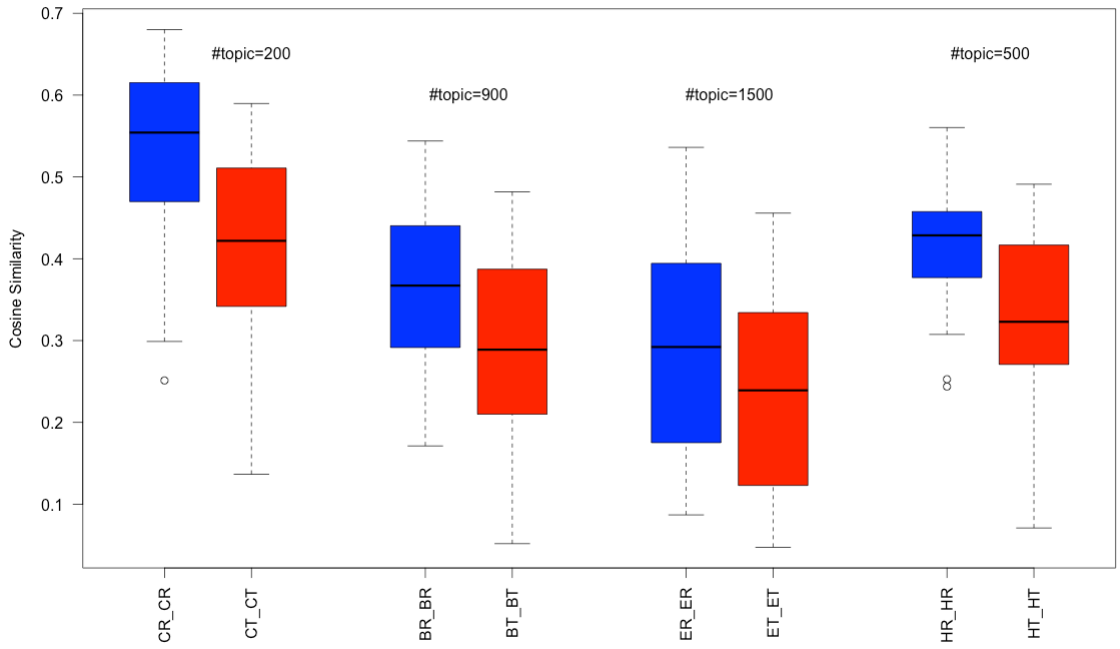


Figure 143: Latent Semantic Similarities between User Profiles and Their 10 Nearest Peers

11.0 STUDY 5: CROSS-SYSTEM USER MODEL TRANSFER FOR RESOLVING COLD START PROBLEMS

In light of frustrating results about the value of external information reported in Chapter 10.0 , this chapter reports another user study focused on the value of external information. In this study, we examine different kinds of external information: user models transferred from a different adaptive system. More specifically, this study compares methods of cross-system user model transfer across two large real-life systems. It investigates the value of transferring user models built for information seeking of scientific articles in the SciNet exploratory search system, operating over tens of millions of articles, on cold-start recommendation of scientific talks in the CoMeT talk management system, operating over hundreds of talks. The user study focuses on transfer of novel explicit *open user models* curated by the user during information seeking. Results show strong improvement in cold-start talk recommendation by transferring open user models, and also reveal why explicit open models work better in cross-domain context than traditional hidden implicit models. In this context, this chapter attempts to re-examine the prospects of the unigram-level user model transfer across related, but different domains. To fight the known problems of unigram profile transfer, a different kind of profiles were explored. While more traditional implicit unigram-level user model is served as a baseline in this chapter, main emphasis is on transfer of an explicit model of interest that is open to the users in the source systems and explicitly curated by them. A subset of results presented in this chapter was published in (Wongchokprasitti et al, 2015).

11.1 USER MODEL TRANSFER

In this chapter, the idea of *user model transfer* to enable warm start in cross-system recommendation scenario was investigated. The idea is "user models can be established in a *source system* and then used in another *target system*." While this idea is not new, past research on user model transfer produced mixed results. This chapter expands earlier research by exploring transferability of *open user models*. A scenario investigated is, users or the source systems have the ability to explore and curate their model by visual interaction. The open user modeling approach is believed to produce better quality user models that could be especially valuable for cross-system transfer. To assess this hypothesis, this chapter shows how the cross-system transfer of open user models improves performance of recommender methods both in the extreme cold-start setting when no preferences are available from the user, and over time when some preferences from the user become available. This chapter shows different transfer strategies in an academic information setting where the source system is a search system for scientific papers and the target system is a system for sharing academic talks. The open user models in the source system are built from visual interaction with keywords and other available information includes views and bookmarks of query results; models in the target system are built from ratings of talks.

This chapter aims to explore the transferability of open user models. Its main contributions are: (a) to show cross-system transfer of open user models greatly improves cold-start recommendation performance, (b) to investigate different ways of transferring open user models from an information seeking system to a talk recommendation system, as well as transfer of more traditional implicit and explicit document information, and show the open user models

bring the greatest benefit, for which an explanation is provided by analysis of cross-system similarities of the different information types.

11.2 SCINET

SciNet (Ruotsalo et al., 2013) is an exploratory search system. SciNet indexes over 50 million scientific documents from Thomson Reuters, ACM, IEEE, and Springer. Going beyond text-based queries, SciNet helps users direct exploratory search by allowing them to interact with an *open user model* discussed below. The approach (called ”interactive intent modeling”) significantly improved users’ information seeking task performance and quality of retrieved information (Routsalo et al., 2013), thus the open user models are promising for cross-system transfer.

Open User Models in SciNet

Unlike traditional information seeking systems, SciNet opens its user model by allowing users to directly see a visual representation of the model and interact with it. User intent is modeled as a vector of interest values over a set of available keywords. The vector can be seen as a bag-of-keywords representation of an “ideal'” document containing keywords in proportion to their interest values. Figure **144** shows the interface: the radial display (A) represents the open user model by showing keywords for interaction; inner keywords (C) represent estimated intent and outer keywords (B) represent alternative intents. The user can inspect keywords with a fisheye lens (D) and bookmark documents. The open user model is visualized as a radial layout where the estimated search intent and alternative intents are represented by scientific keywords,

organized so that keywords relevant to the user are close to the center and similar intents have similar angles. Users start by typing a query and receive a list of documents, which they can bookmark, and then direct the search by interacting with the user model. Users can inspect model keywords shown on the radar and *curate* the model by dragging keywords to change their importance. After each iteration, the user model is inferred from the whole set of user actions, documents are searched based on the updated model, and the radial visualization is updated.

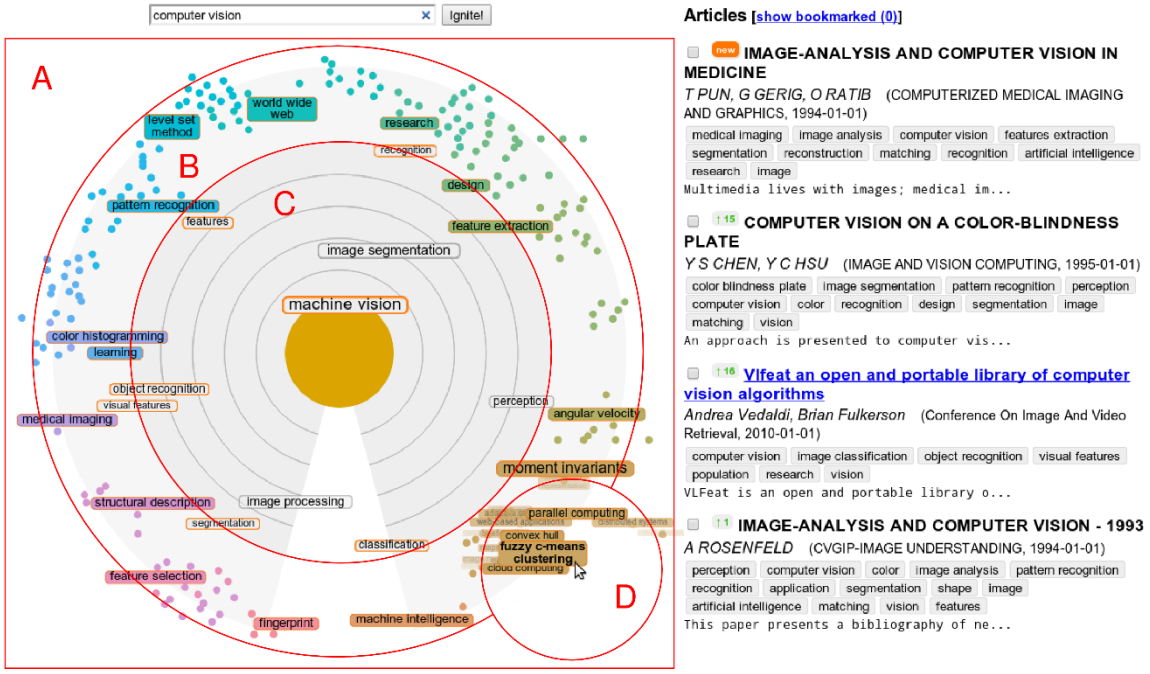


Figure 144: The SciNet System

The primary interest is to use (1) the whole content of the open user model and (2) its curated subset (the keywords that the user moved in the process of curation). As a baseline, the transfer of more traditional information is explored: (3) the set of relevant documents selected by the user in the process of search (which could be considered as hidden, implicit user model) and

(4) a broader set of all documents retrieved in response to user queries that is a weaker reflection of user interests.

11.3 MODEL TRANSFER FOR CROSS-SYSTEM RECOMMENDATION

The goal of this cross-system setup is to recommend users relevant talks based on features extracted from the description of the talk and their model of interests. Each CoMeT talk is represented by a unigram model, that is, as a vector of word counts in the description of the talk, normalized by the maximum word count, and converted into term frequency–inverse document frequency (TF-IDF) representation. Transferring user models from SciNet creates a warm start for CoMeT talk recommendation. Two approaches to transfer the explicit open user model from SciNet, and two approaches to transfer implicit user models from the SciNet search trace were investigated.

"Baseline": Ranking based on the CoMeT user model alone

The baseline method ignores SciNet, and ranking is based only on the CoMeT user model. In the pure cold-start case where no bookmarks are available, the baseline is unable to give recommendations.

"MA.Keywords": Transfer the explicit open user model from manipulated keywords

The SciNet open user model represents the user's interest over keywords. While the model predicts importance over all keywords, a subset of most promising ones are shown for interaction, and a further subset out of those manipulated by the user. The set of all scientific

keywords dragged by the user on the SciNet interface during the search session is treated as a pseudo-document containing the keywords is taken into account. Each keyword is associated with a weight corresponding to the user's interest (radius where the user dragged the keyword). The pseudo-document is converted into a vector of unigrams by taking each unigram within each keyword (e.g. "support" and "vector" within "support vector"), associating it with the corresponding weight of the keyword (or sum of weights if the unigram occurs in several keywords), and discarding unigrams that do not appear in the CoMeT corpus. This extracts from the SciNet open user model the unigram information that is matching the CoMeT information space. Since the open user model is represented as a single pseudo-document instead of a corpus, the corresponding unigram vector is not converted into TF-IDF representation. Instead, it is only normalized by its maximum value. As a result, the corpus of CoMeT talks is then converted into TF-IDF over CoMeT talks only. The resulting unigram vector of the pseudo-document is added into the user's cold-start set of bookmarked CoMeT talks.

"SH.Keywords": Transfer the explicit open user model from shown keywords

The SciNet open user model displays the subset of most important keywords to the user at each of search iterations (area A in Figure **144**). For each of search iterations, the keyword subset is taken as a pseudo-document, where each keyword is associated with a weight corresponding to the user interest predicted by SciNet. Each such pseudo-document is converted into a vector of unigrams in the same way as in Subsection 0. As a result the user gets one vector of unigrams for each of search iterations, which together represents the evolution of the user model over the search session. As in Subsection 0 this is not a corpus, hence each vector of unigrams is normalized by its maximum value and CoMeT talks apply TF-IDF weighting over

CoMeT talks only. The unigram vectors of the pseudo-document are added into the user's cold-start set of bookmarked CoMeT talks.

"BM.Papers": Transfer an implicit user model from bookmarked documents

Scientific documents bookmarked by the user during the SciNet search session provide implicit information about the user's interests. All bookmarked documents were converted as the same unigram representation as CoMeT talks as in Subsection 0. Since the bookmarked documents form a corpus, unigram vectors of CoMeT talks and SciNet documents were converted into a TF-IDF representation computed over both corpuses. For each user, the resulting unigram vectors of the SciNet bookmarks were added into the cold-start set of that user's bookmarked CoMeT talks.

"SH.Papers": Transfer an implicit user model from shown documents

The method conducts the same as in Subsection 0 but using all documents seen by the user during the SciNet search session (not only bookmarked documents).

11.4 USER STUDY FOR DATA COLLECTION

To perform experiments on cross-system model transfer, data were collected to capture interests of the same users in the two explored systems. This data was collected through a user study In Section 11.5, experiments were performed on cross-system transfer from SciNet to CoMeT, for cold-start prediction of CoMeT talk attendance using information from the SciNet search trace. Before performing the recommendation experiments, a task-based user study in a laboratory setting was conducted to collect data for a set of users over both systems:

1. The search trace and open models of the users were collected when they conducted an exploratory search in SciNet for scientific literature corresponding to their research interests. Participants were asked to imagine that they are preparing for a course or seminar on their research interest.
2. User preferences in attending academic talks indexed in the CoMeT system were collected. Participants were asked to bookmark interesting talks and rate to what extent they would like to attend it.

11.4.1 Task Descriptions

For SciNet, a search task was chosen to be complex enough that users must interact with the system to gain the information needed to accomplish the task, and broad enough to reveal research interests of users. The task is: “Write down three areas of your scientific research interests. Imagine that you are preparing for a course or a seminar for each research interest. Search scientific documents that you find useful for preparing for the courses or seminars.” To determine which documents were relevant to the task, the users were asked to bookmark at least five documents for each research interest. For CoMeT, the following rating task was used: “Please rate all the scientific talks whether you would like to attend the talks or not. If you don’t want to attend the talk then just click “no” button and go to next talk. If you want to attend the talk then you click “yes” button and fill the ratings.” The following guidelines for ratings were provided:

“5” : This talk matches my research interest and I would definitely attend it.

“4” : This talk matches my research interest and I would likely attend it.

“3” : This talk somewhat matches my research interest and I might attend it.

“2” : This talk somewhat matches my interest, but its unlikely that I attend it.

“1” : This talk somewhat matches the research interest but I wouldn’t attend it.

11.4.2 Participants

Twenty researchers (fourteen male and six female) from University of Helsinki were recruited to participate in the study. All participants were research staff (ten PhD researchers and ten research assistants) in computer science or related fields. The participation was limited to researchers because the nature of SciNet and CoMeT required participants having experience in scientific document search and having interest in attending research related talks or seminars. Prior to the experiment, a background survey of the participants was conducted to ensure that they have conducted literature search before and have also attended research related talks or seminars.

11.4.3 Procedure

The study used a within-subject design in which all the participants performed both the tasks using both the systems alternatively. To minimize the impact of experience in one system on another, the order of system use was counter balanced. Ten participants used the SciNet system first and then used the Comet system while the remaining ten participants used the CoMeT system first and then used the SciNet system. The protocol for each system had two stages: for SciNet the stages were demonstrating the system (7 minutes) and then performing the search task

by the participant (30 minutes); for CoMeT the stages were demonstrating the system (7 minutes) and then performing the rating task by the participant (75 minutes).

11.4.4 Data Logging

When the participants were performing the search task in SciNet, all their interactions with the system Data were logged from each interaction included details of the scientific documents displayed, keywords for the estimated intent predicted by the system and for alternative intents (areas C and B in Figure **144**), curated keywords, search query of the users, abstracts of scientific documents viewed, scientific documents bookmarked by the users and the corresponding timestamps of all the user interactions. When the participants were performing the rating task in CoMeT, the ratings of the scientific talks or seminars and the ratings of the novelty of the scientific talks or seminars were logged.

11.5 CROSS-SYSTEM RECOMMENDATION EXPERIMENT

The user data gathered in Section 11.4 was used to create experiments on cross-system recommendation of CoMeT talks by transporting user models from Scinet to CoMeT. Four modes of transporting user models were compared and described in Section 11.3. Both the global impact in a setting with several rated talks available from each user, and performance in the harder cold-start setting was evaluated.

11.5.1 Data Processing and Demographics

There were 500 CoMeT talks selected from January 10 to February 5, 2013, containing 8,406 unique unigram terms. SciNet indexes over 50 million scientific articles, out of which users see a subset based on their interest and search behavior: there were 9,457 unique SciNet articles returned to our participants with 30,848 unique terms, of which 5492 of them or 17.80% overlapped with CoMeT corpus. SciNet also records keywords shown to the user and manipulated by the user; in total the participants were shown 3,474 unique keywords (1974 of them or 56.82% overlapped) and manipulated 178 unique keywords (157 of them or 88.20% overlapped). Scientific documents and keywords from SciNet and research talks (talk descriptions) from CoMeT were cleaned from html tags, stop words were removed, and stemmed by the Krovetz algorithm. In each cross-system transfer approach, the documents, talks, and keywords were converted into unigram vectors with according to term frequency–inverse document frequency (TF-IDF) schemes, as described in Section 11.3.

11.5.2 Experiment 1: Global Impact of Cross-System Models

In the first study, the traditional (non-cold-start) learning setting was taken into account where much training data is available within CoMeT, that is, recommendation for users that have used both CoMeT and SciNet for some time. This evaluation was called “global impact of the transfer”. The CoMeT-only baseline was compared against four transfer approaches: explicit open user model from manipulated or shown keywords, and implicit user model from bookmarked or shown documents. For each approach, two methods from the previous studies

(Centroid and positive-sample-only k-Nearest Neighbors) and new one (k-Nearest Neighbors) were used to make recommendations based on the transported user model.

– **Centroid**: the centroid (mean) of the unigram vectors of the user's bookmarked CoMeT talks and any vectors transferred from SciNet was taken into account. Test talks are ranked by cosine similarity of their unigram vector to the centroid.

– **k-Nearest Neighbors (KNN)**: unigram vectors of bookmarked CoMeT talks and vectors transferred from SciNet were treated as positive samples, and vectors of non-bookmarked CoMeT talks as negative samples. For each test talk, its *k* neighbors (nearest positive or negative samples) were detected by cosine similarity of unigram vectors. Test talks are then ranked by $S_{pos} - S_{neg}$, where S_{pos} is the sum of cosine similarities from the test talk to the positive neighbors and S_{neg} is the sum of cosine similarities to the negative neighbors.

– **Positive-Sample-Only k-Nearest Neighbors (denoted KNN.PO)**: this method is similar to a KNN approach but it only considers the k neighbors of a test talk from the positive samples only, and then ranks the test talk by the sum of cosine similarities from the test talk to the *k* positive neighbors (Schwab et al., 2000).

11.5.2.1 Experimental Setup: Ten-Fold Cross-Validation

Ten-fold cross-validation was setup. Bookmarked and non-bookmarked CoMeT talks of each user were randomly divided into ten equal-size bins, and evaluation was run ten times with each bin in turn held as the test set and the other nine bins as the learning set used to construct user models. SciNet user model data was transferred to each of learning sets by the transfer approaches to augment the resulting user model. Results (rankings of test talks) are evaluated by

mean average precision (MAP): mean of precision values at locations of positive test talks in the ranking, averaged over users and cross-validation folds.

11.5.2.2 Results of experiment 1

Table 7 shows the results of global transfer models from SciNet data to CoMeT system. Performance of the CoMeT-only baseline varies by recommendation approach: Centroid and KNN perform comparably and KNN.PO is slightly better. For the transfer approaches centroid performs comparably to other approaches; KNN or KNN.PO can yield slightly higher MAP but depend on value of k. Overall, in this non-cold-start setting transfer of SciNet information did not yield much improvement over the CoMeT-only baseline: transferring an implicit model from shown SciNet documents underperformed the CoMeT-only baseline and the other transfer approaches were not significantly different from that baseline. In this non-cold-start situation user profiles in CoMeT had enough data to work well on their own. As a result, a greatest benefit of transfer in the cold-start setting is expected in the next study.

Table 7: Global Impact MAP from Transferring Models by Transporting Methods, and by Algorithms

MAP	**Centroid**	**5NN**	**10NN**	**20NN**	**30NN**	**5NN.PO**	**10NN.PO**	**20NN.PO**	**30NN.PO**
Baselines	0.47	0.45	0.47	0.48	0.46	0.48	0.50	0.50	0.50
BM.Papers	0.42	0.44	0.45	0.45	0.44	0.43	0.44	0.44	0.44
SH.Papers	0.36	0.36	0.36	0.35	0.35	0.36	0.37	0.36	0.36
MA.Keywords	0.48	0.46	0.48	0.48	0.47	0.49	**0.51**	**0.51**	0.50
SH.Keywords	0.47	0.46	0.48	0.49	0.49	0.48	0.49	0.49	0.48

This study focuses on the cold-start settings. Again, there are four kinds of user models from SciNet (explicit open user models from shown and manipulated keywords, and implicit models from shown and bookmarked documents).

11.5.2.3 Experimental setup: Ten-Round-Ten-Fold Cross-Validation

A range of cold-start was setup where the user has bookmarked 0-20 CoMeT talks. For each number of bookmarked talks, a ten-fold cross-validation was taken into account. Data was divided into ten bins. In each fold, one cross-validation bin was held out as test data, and from the remaining nine bins, a pool of bookmarked talks and a pool of non-bookmarked talks were sampled. Within each fold, random sampling was performed 10 times (10 “rounds”) and report average results over rounds. For each number of bookmarked talks, the number of sampled non-bookmarked talks was chosen to keep the same ratio of bookmarked to non-bookmarked talks as overall in the data of that user. The evaluation results were reported by the same MAP criterion as in the first experiment.

11.5.2.4 Results of Cold-Start Recommendation

The cold-start recommendation for the CoMeT-only baseline using the three recommendation methods (centroid, KNN, KNN.PO as in Section 11.5.2) was tested. Centroid and KNN.PO models performed equally (results omitted for brevity; essentially no visible difference regardless of the number of bookmarks) and outperformed the KNN models. Given that the Centroid model is simpler and faster than KNN.PO, it was used for all transfer methods in the cold-start recommendation. Figure **145** shows the results of transfer in the cold-start setting. Explicit transfer of the SciNet open user models (ma.keywords and sh.keywords) helped smooth

the transition of the new users into the CoMeT system. In fact, the MAP results of keyword models are significantly better than the baseline until the user has two bookmarked talks in the SciNet implicit keyword transfer centroid model, and until five bookmarked talks in the SciNet explicit keyword transfer centroid model. While the transfer of explicit curated models worked well, the transfer of implicit models built from retrieved and bookmarked papers harmed the MAP performance. Even the model built from explicitly selected talks performed poorly comparing to the baseline. The analysis to reason for the much better performance of explicit user models is in the next section. Figure **145** explained that the MAP was shown for each method and each number of bookmarked CoMeT talks. Error bars showed 95% confidence interval of the mean over users and cross-validation folds. "20 users"–"18 users" denoted how many users had enough data within cross-validation folds to have the desired number of bookmarked talks.

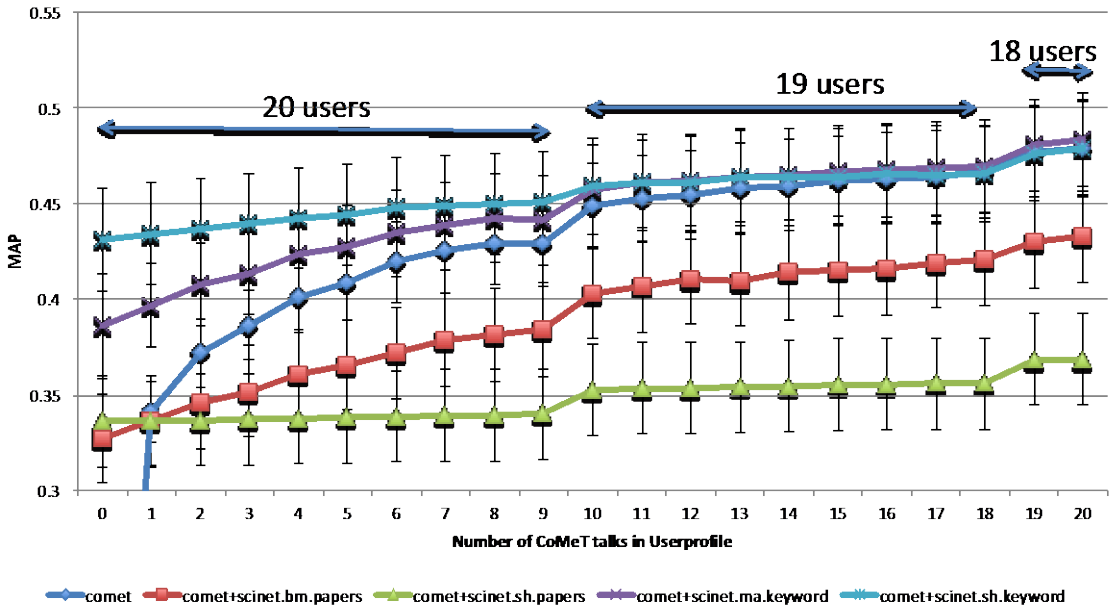


Figure 145: Cold-Start-Effect MAP Results of Centroid Models

11.6 SUMMARY AND DISCUSSION

From the result, the transfer of explicitly curated user models yielded better results than alternative transfer approaches and the baseline. The reason for the good performance of open curated models can be analyzed by examining how similar the different types of transferred information to the information contained in the bookmarked and non-bookmarked CoMeT talks.

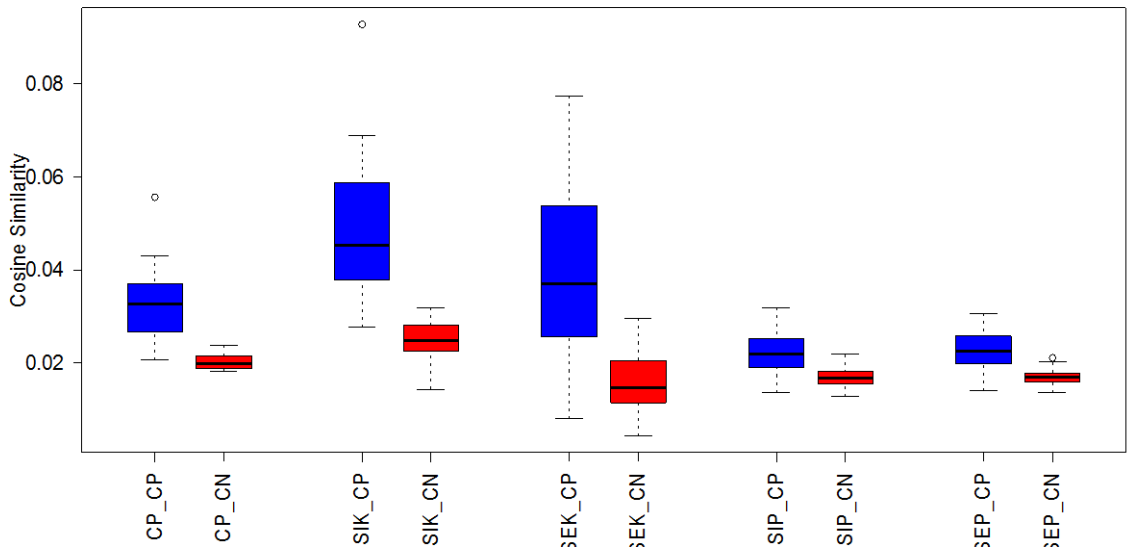


Figure 146: Distributions of Similarities between CoMeT Talks and Five Representations of User Interests

The results of this analysis are shown in Figure **146**, which plots the distribution across users of average similarities from five representations of user interests (bookmarked talks in CoMeT and four kinds of models transferred from SciNet) to bookmarked/non-bookmarked talks in CoMeT. As expected, bookmarked talks in CoMeT are closer to each other (CP_CP boxplot) than to non-bookmarked ones (CP_CN boxplot). More interesting is that both manipulated

SciNet keywords (SEK_CP boxplot) and shown SciNet keywords (SIK_CP boxplots) are even closer to the space of bookmarked talks, separating them quite well from non-bookmarked ones (SEK_CN and SIK_CN boxplots, respectively). As a result, incorporating open curated keywords into the Centroid model helped alleviate the cold-start problem and improved recommendation performance. In contrast, implicit models built from shown (retrieved) SciNet papers and bookmarked SciNet papers (SIP_CP and SEP_CP, respectively) are quite far from the space of bookmarked talks and offer poor separation of this space from the non-bookmarked talks (SEP_CN and SIP_CN, respectively). Consequently, implicit models add more noise than value and damage recommendation.

Overall, this chapter explores a novel approach to cross-system personalization based on transferring an explicit, open, and editable user model maintained by one system to another system in a similar, yet different domain. More specifically, this chapter explored whether an open user model maintained by SciNet exploratory literature search system could help recommend relevant research talks to users of CoMeT talk sharing system who have established SciNet interest models. The impact of open model transfer was compared with a baseline case that uses no information from SciNet and with more traditional implicit model transfer based on information about retrieved and marked papers. The comparison examined an overall impact of the transfer as well as its value in a cold-start situation when the user has none or too few talks bookmarked in CoMeT. While the first study not able to register a significant impact of model transfer in a general case, the second study demonstrated a significant positive impact of the open model transfer in the cold-start case. It also demonstrated that the use of open, explicitly curated user models is critical for the success of user model transfer: the transfer of SciNet data in the form of implicit model damaged the performance of talk recommendation. The analysis of

differences between explicit and implicit model hinted that the use of explicitly curated models reduces the noise in modeling user interests. An interesting research question for the future work is to what extent machine learning approaches could simulate model curation focusing on better representation of user interests. Several feature selection approaches were planned to explore and compare them with the impact of manual curation.

12.0 ANALYSIS AND DISCUSSION FOR ALL STUDIES

Results of all studies were analyzed in detail by design space analysis and dataset perspectives. Design space consists of five different factors that were explored to find the optimum models. The dataset perspective is the last section. Results are discussed in a series of sequential datasets based on the chronological progression of the studies.

12.1 DESIGN SPACE ANALYSIS AND DISCUSSION

In this section, multiple factors are compared and explored in our design space across all studies in order to examine the observed differences and discuss possible reasons for these differences. Section 12.1.1 examines the role of different types of external sources. Section 12.1.2 investigates the effect of different user profile representations. Section 12.1.3 looks at the role of recommender types. Section 12.1.4 explains the effects of external unigram terms exclusion, and Section 12.1.5 shows the analysis of recommendation fusions.

12.1.1 TYPES OF EXTERNAL SOURCES

12.1.1.1 Personal Webpage

In Study 1 of the CN3 dataset, 117 users with homepage sources had a median of four pages. Many homepage-augmented models with different kinds of settings among 317 CN3 users were explored. Among them, the homepage-augmented CBF model with 20-NN.PO CN3-term SVD-300-latent-topic setting was selected as a CBF representative, and the homepage-augmented CBCF one with all-term SVD-900-latent-topic setting was selected for a CBCF representative, because they both yielded the highest MAP results as compared to other models. The results from Study 1 showed that models with homepage augmentation provided some improvement in recommendations when compared with baseline models but the difference was not significant. The results were validated with the external validity of the holdout data on the CN3 dataset. The validation showed the similar results. To evaluate the effect of external source augmentation on the lab study with real users in study 4 of the CoMeT dataset, 44 participants were recruited. Of the 44 participants, 29 provided their homepages. They also had a median of four pages, similar to CN3 users. The results showed the homepage-augmented models of both CBF and CBCF recommender types did not improve the recommendation using the MAP measures, or nDCG measurements on Relevancy, and Novelty. In summary, the homepage source neither harmed nor helped to improve recommendation in the global impact scenario of either CN3 or CoMeT datasets.

While in the general situation, the homepage source did not provide any impact on the recommendation in the CN3 study; however, it did help alleviate the severity of the cold-start problem. Results from Study 2 showed that the homepage-augmented CBF model significantly

improved the recommendation in the early cold-start stages—from no-bookmarked-CN3-talk to two-bookmarked-talk windows. The homepage-augmented CBCF model provided significant impact only in the first stage of the cold-start scenario when a bookmarked CN3 talk did not exist in the user profile. The external-validity study of cold-start problems showed similar results—improvement on both CBF and CBCF recommender types. In the CoMeT user cold-start study, the homepage-augmented CBF and CBCF models provided significant improvement only in the first stage when there were no bookmarked CoMeT talks in the user profile. From these results, recommending models with homepage augmentation gained significant positive impact in the early stages of cold-start scenarios.

12.1.1.2 Bibliography

In the external-source-augmented recommendation improvement study using the CN3 dataset, the bibliography-augmented 5-NN.PO with CN3-term full-text setting was selected as a CBF representative. The bibliography CBF model produced significant improvement in recommendations over baseline. While the bibliography CBF model worked well, the bibliography-enhanced CBCF model did not provide any improvement. In the latter study of cold-start problems, bibliography-enhanced models from both CBF and CBCF recommender types performed significantly better than the baselines in the beginning-stage cold-start situation when there were no bookmarked talks in the user profile. Only bibliography CBF models were able to significantly outperform the CBF baseline. Two hypotheses could explain the significant difference in results between CBF and CBCF models. First hypothesis is that the CN3 dataset containing partial user bookmark ground truth may have adversely affected the performance of the CBF models. Another hypothesis is that results using CBCF was artifactual. By artifact, we

mean that users who had similar or close research interests tended to attend the same conferences and were unlikely to attend other unrelated conferences. As a result, the potential talks to be recommended from the CBCF, which were gathered from bookmarked talks of the nearest peers, excluded ones that were unrelated or uninteresting. Also, the number of users in the CN3 study was considerably high compared to the CoMeT user study. With a larger number of users to be considered, nearest peers chosen were more likely to be better quality.

In contrast with studies with CN3 dataset, bibliography-enhanced models in the CoMeT user studies did not perform in the global-impact and cold-start studies. The bibliography-augmented 10-NN.PO with full-text setting produced a comparable MAP result to the baseline CBF in the global-impact study, as did the bibliography-augmented CBCF model with SVD-900-latent-topic settings. In the latter cold-start problem study, bibliography-enhanced models from both CBF and CBCF recommender types only performed significantly better than baselines in the *beginning-stage* cold-start stage. The difference between CBF and CBCF models was narrow as a result of the CoMeT dataset containing complete user-bookmark ground truths, making MAP results of CBF models comparable to CBCF counterparts. Also, a number of users selected as nearest peers were much smaller than the CN3 studies due to a large number of different users from both datasets (CN3: 815 and CoMeT: 44). Another observation was that a median of user publications was 27 in the CN3 study, which was apparently higher than six publications in the CoMeT studies. This might be the key point. It was already known that a small bibliography is not as helpful. However, the CoMeT study dealt with starting researchers, who had very few publications. These are the kind of people for whom CN3 bibliography has not been working well.

12.1.1.3 External Bookmarked Scholarly Papers

Only a small number of users provided the external bookmarked scholarly paper sources in both the CN3 and CoMeT studies. There were 45 for CN3 and 34 for CoMeT. The median number of papers from CN3 users was 72, which was higher than the 30 papers from CoMeT users. In the preliminary study presented in chapter three, we explored user tags from the CiteULike source, which was one of the external bookmarked scholarly paper sources. The preliminary results showed that adding tags into the vector space CoMeT model helped improve novelty and relevance of recommendations. Hence, this source was expected to provide positive impacts in this thesis. Surprisingly, results showed that adding the content of external bookmarked papers into the CBF and CBCF user model in the global-impact study on both CN3 and CoMeT datasets did not improve recommendation performance. In the cold-start problem study, the external-bookmark-enhanced CBF and CBCF models statistically outperformed baselines only in the very early cold-start stage, where there were no bookmarked talks in the user profile of both CN3 and CoMeT datasets. In the latter cold-start stages, the external-bookmark-augmented CBF and CBCF models from both CN3 and CoMeT datasets did not improve recommendation compared to baselines. Moreover, the external-bookmark-enhanced CBF model in the cold-start CN3 study significantly underperformed the baseline during 1-talk to 3-talk *beginning-stages*.

When examining unigram terms from content of base corpus and external bookmark counterpart, around half of external-bookmark unigram terms did not appear in the CN3 corpus or CoMeT one. The content of this external source formed a considerably different term space than the base corpus.

12.1.1.4 SciNet User Profiles and Search Logs

CoMeT studies were meant to be retrospective studies of the previous CN3 studies. However, with frustrating results, cross-system CoMeT-SciNet studies became retrospective, as well. The goal was to find other useful external sources as a result of the previous failure to find good external sources. However, because there were only twenty participants in the dataset, the CBCF models were dropped from these studies. When measuring the global impact on external data, cross-system study results showed similar trends to the CoMeT study. For example, there was no significant positive improvement for CBF models with neither explicit nor implicit SciNet user models. Four transfer approaches performed comparably to the baselines. One hypothesis was that the user profiles had enough user information from bookmarked CoMeT talks. Adding extra information did not provide any impact on the recommendation performance.

However, "MA.Keywords" and "SH.Keywords" CBF models with SciNet user models provided a positive impact on the cold-start recommendation study. They helped significantly improve the recommendation performance beyond the baseline from the *beginning-stage* cold-start situation until the two-bookmarked-talk stage in the SciNet implicit keyword transfer centroid model and until the five-bookmarked-talk stage in the SciNet explicit keyword centroid model. In fact, explicit transfer of the SciNet open user models ("MA.Keywords" and "SH.Keywords") considerably helped smooth the transition of the new users into the CoMeT system. One hypothesis was that adding a few hundred relevant terms or keywords into the user profiles, when there was little or no information about users, provided a significant impact on the recommendation performance.

12.1.2 USER PROFILING

12.1.2.1 User Profile Representation

In this dissertation, two different representations of user interests, classic unigram profiles and LSI-based profiles, were compared. The LSI-based approach aims to take advantage of latent semantic analysis of the dataset. In general, user models with simple unigram representations of user interests performed consistently well with any recommender type or any external source. LSI-based profiles yielded better results, but they need to be tuned. In study 1 of the CN3 dataset, nine out of twelve representative user models were LSI-based representations of user interests. In the global-impact study of the CoMeT dataset, half were classic unigram representations of user interests. Finally, in the cross-system CoMeT-SciNet study, only unigram representations of user interests were considered. In conclusion, classic unigram profile representation is considered a golden standard for user profile construction, but LSI-based representations of user interests might not always surpass unigram representations of user interests.

12.1.2.2 User Profile Granularity

One study goal was to explore how a transition from the individual user profile to the cluster profile is able to tackle data sparseness and cold-start problems. The idea of the cluster representation of user interests is to group the individual profile with ones of their peers. However, the CN3 study demonstrated that the group-level models with classic unigram representations of user interests yielded poor results. No cluster model was selected as a representative model for any external source augmentation or baseline. Apparently, cluster

models in all three clustering methods with the space dimension (unigram profiling vs. LSI-based profiling) or feature selection (considering all terms from CN3 and extra terms introduced from external sources or excluding only target terms from CN3 corpus) underperformed other external-bookmark-augmented centroid or KNN.PO models. Even though applying LSI-based user representations of user interests to the group-level models increased performance, it was still significantly below the individual models. One hypothesis was that the group-level representations of user interests not only increased relevant unigrams but also considerably increased noise or irrelevant terms in the profile. As a result, the group-level user model was excluded from CoMeT and cross-system CoMeT-SciNet studies.

12.1.2.3 User Profile Application

Another component of the explored design space was application of user profile for personalization. The single-interest (Centroid) user models and multiple-point-interest counterparts were investigated and compared. The multiple-point-interest user models were constructed by two methods, the *k*-Nearest-Neighbors (KNN) and positive-sample-only *k*-Nearest-Neighbors (KNN.PO). These user profiles were applicable only to the CBF recommender type.

The KNN was investigated only in the cross-system CoMeT-SciNet study and did not show positive improvement when negative samples (non-bookmarked talks) were added into the profiles. As a result, the only available performance comparison was a number of applications selected between Centroid and KNN.PO in the CN3 and CoMeT studies. The Centroid vector space models have been recognized as standard in the information retrieval community. The Centroid models were among top performers in every CBF study. Four centroid models out of

six representatives, with or without external source augmentation, were chosen in the CN3 study. Even though none were in CoMeT, their results were close to the top.

In the latter multi-point-interest applications, two KNN.PO were selected in the CN3 study, and all six representatives were KNN.PO. One characteristic of KNN.PO is that the KNN.PO models behave the same way as the Centroids when a number of bookmarked talks in the user profiles are less than designated *k* closet peers. Once the profiles grow beyond *k*, KNN.PO takes advantage of *k*-Nearest-Neighbors algorithm that preserves multi-user interests.

One observation was that centroid applications dominated in the CN3 study, but KNN.PO counterparts took over in CoMeT. The hypothesis was that the CN3 dataset contained only partial ground truth of user interests while CoMeT dataset stored all of them. Consequently, the partial and incomplete CN3 user profiles worked well with the centroid application, except when adding extra user information from homepage or bibliography sources.

Augmented user profiles with bibliography worked really well with KNN.PO with *K* being only 5. With small *K*, most of KNN.PO models showed the sign that they had more than one user-interest point. In the CoMeT study, once the user profiles had an ample size of bookmarked talks, the KNN.PO applications worked well. All of selected CoMeT representative models were KNN.PO applications. Also, most of them resulted in small *K*, which were ten or less. It was also observed that with complete ground truth in CoMeT corpus, baseline KNN.PO models produced very high MAP results, compared to ones in the CN3 study. Adding extra information did not provide any significant impact.

12.1.3 RECOMMENDER TYPES

Recommender types explored in this dissertation were content-based filtering (CBF) and content-boosted collaborative filtering (CBCF). The CBF was explored in all five studies and one external-validity study. The CBCF was investigated in all studies except the cross-system SciNet-CoMeT study, because a number of users in the dataset was too small to generate reliable results for CBCF.

As shown in chapter six, the CBF models delivered very low MAP results in the CN3 study but relatively high results in the CoMeT study. As mentioned before, we hypothesize that the outcomes were a result of the experimental design in the CN3 cross-validation study. The CN3 study design treated all non-bookmarked talks as negative samples (assuming that not bookmarking a talk was an implicit classification of it as not relevant), even talks for conferences that users did not attend, and thus had no real chance to view and judge the talks. Coupled with partial ground truth of the CN3 dataset, MAP results from CBF recommender models were around 0.06. On the other hand, in the CoMeT study, MAP results of CBF models were around 0.36. In the CoMeT study, no assumptions were made about negative feedback, because all participants needed to provide explicit feedback for all talks in the experiments. From another perspective of partial ground truth on the CN3 dataset, the user profile did not have enough samples to reliably represent interests of users. Adding useful and relevant information, such as bibliography or homepage segments, into the user profiles provided CN3 CBF models with enough samples. As a result, those CN3 CBF models produced a positive impact on recommendation performance. In particular, CN3 CBF model with bibliography augmentation yielded significantly better MAP results than the CBF baseline counterpart. Additionally, CN3

CBF models with either homepage or bibliography augmentation significantly helped smooth the cold-start problem in recommendation. However, CoMeT CBF models produced high MAP results. One hypothesis was that CoMeT user profiles contained enough samples due to complete ground truth collected in the CoMeT dataset. Even though CBF models with explicit user model augmentations in the cross-system CoMeT-SciNet study did provide slightly better results, the difference was not significant. Consequently, adding extra user information from external sources did not provide any significant impact.

In another explored recommender type, CBCF, MAP results had similar levels across both CN3 and CoMeT studies. Unlike CBF results, CN3 results were relatively low compared to CoMeT, which were apparently higher. In the CBCF, MAP results from the CN3 study were around 0.23 and 0.40 while CoMeT study MAP results were around 0.40. Even with having partial samples in the user profiles in the CN3 study, results were still much higher than ones from CBF counterparts. One hypothesis was that each conference was attended by peers who had considerably similar interests. As a result, potential talks to be recommended were from their peers' bookmarked talks. Those bookmarked talks were more likely to be from conferences that users had attended rather than the conferences users had not. This effect increased the chance that the correct bookmarked talks would be recommended by eliminating talks from the conferences users were unlikely to attend from the potential ones. By removing those talks from the potential talks, CBCF recommenders were more robust to the experimental design effect than CBF counterparts, or CBCF models were regarded as an artifact of this study. The cold-start problem study from both CN3 and CoMeT datasets yielded consistent results. The external-source-augmented CBCF models provided a significant impact only in the beginning stage where there was no bookmarked talk in the user profiles. But after the *beginning-stage* of a cold start

situation, there was no significant positive improvement. Results show that injecting extra user information into the profiles did not provide any impact on CBCF recommendations. The CBCF models, with or without external source augmentation, seemed to be able to generate a similar level of performance after the user profiles had one bookmarked talk.

12.1.4 OUT-OF-CORPUS UNIGRAM TERMS EXCLUSION

In classical bag-of-words vector space representations of user interests or LSI-based, two sets of unigram terms were explored, one from base corpus only and another from base corpus combined with the augmented external source.

Table 8: The Number of Unigram Terms in Each Source and Overlapping Percentage Compared to Base Corpus

	#Terms	% in CN3	#Terms	% in CoMeT
Homepage	17,137	29.29%	4,928	54.14%
Bibliography	7,888	60.92%	3,532	71.15%
Ext. Bookmark	12,093	50.52%	8,744	49.35%
SciNet.Papers	N/A	N/A	30,848	17.80%
SciNet.Keywords	N/A	N/A	3,474	56.82%

For the CBF recommender type, as shown by the results from both CN3 and CoMeT studies, CBF models with outside-corpus-term-excluded user profiles performed better than ones that included extra unigram terms under similar recommending configurations. As explained in

the summary subsection in chapter six, target talks contain only terms appearing in the base corpus. Therefore, the extra terms did not increase similarity between either a vector space from the Centroid models or from the KNN or KNN.PO models and target talks. Furthermore, the extra terms decreased the similarity between bookmarked talks in the user profile and target talks by diluting the IDF value, "where-about", of terms in the target talks. Since IDF is a part of similarity measure, similarity between bookmarked talks and target talks are lesser so it is harder to distinguish between interested talks and non-interested ones.

Another plausible hypothesis was that the external sources, which did not have many terms in common with the base corpus, tended to perform poorly. Table **8** shows a total number of unigram terms from each external source on each study. The bibliography, having highest percentage of overlapped terms in CN3 and CoMeT datasets, was a source that significantly boosted the performance of CBF models in the CN3 global-impact setting study and was the best one to smooth the cold-start problem in the CoMeT cold-start study. Also, another similar trend came from the cross-study SciNet-CoMeT cold-start study that transferred CBF models from explicit user models, "MA.Keywords" and "SH.Keywords", which consisted of a higher percentage of overlapped unigram terms with base corpus. They performed slightly higher than ones of implicit user models, "BM.Papers" and "SH.Papers" in the global-impact CoMeT-SciNet study. Moreover, transfer CBF models from explicit open user models; "MA.Keywords" and "SH.Keywords", performed significantly better in the cold-start context than baselines.

Surprisingly, the negative impact of augmenting extra out-of-corpus unigram terms into the user profile did not affect the CBCF recommenders. The CN3 study showed that CBCF models with and without extra unigram terms produced comparable results. There was one plausible hypothesis to explain this result. The CBCF was an artifact of a study. Users that attended the

same conference series were more likely to have similar research interests. They tended to bookmark talks that had unigrams in the same clusters. Therefore, the baseline model performed very well in finding the nearest peers. As a result, adding extra terms from the external source, whether excluding out-of-corpus terms or not, did not impact the CBCF models.

12.1.5 FUSING DIFFERENT RECOMMENDATION APPROACHES

After exploring recommending models with external source augmentation, we investigated the idea of fusing their results to find the good combination of models. In the CN3 study, almost all cross-source fusion models, on both CombMNZ and CombSUM methods, yielded significantly better results than the CBF baselines, but only nine were able to significantly beat the CBCF baselines. In the CoMeT study, none produced improvement over the CBF baselines (the difference was not significant), and none produced any improvement over the CBCF baselines. These results can be explained by the hypothesis that CBF and CBCF baselines had comparably high results. Neither the CBF nor CBCF models outperformed baselines. Therefore, combining them did not yield any significant improvement.

In respect to two recommendation-fusing methods, CombSUM and CombMNZ, the results from the CN3 study indicated that CombMNZ fusion models were more reliable than CombSUM fusion models. CombMNZ weighed target items higher if they were recommended by more recommendation sources while CombSUM did not consider this factor. Also, in the CoMeT study, CombMNZ methods performed slightly better than CombSUM, but the differences were not as obvious as ones from the CN3 fusion study.

12.2 DATASET DISCUSSION

12.2.1 CN3 DATASET

One characteristic of the CN3 dataset was its sparseness. Users did not provide judgments for every single talk in the corpus. They only bookmarked some interesting talks (not bookmarked talks might be still relevant but never judged) and only for conferences they attended. In other words, the dataset contained only the partial ground-truth in the form of user bookmarks. However, the experimental design implicitly assumed that only talks that users bookmarked were of interest and treated the rest of talks as negative samples. This strong assumption affected MAP results, making them lower than an evaluation with the assumption that included only talks that users had attended.

There was only one CBF model, bibliography augmentation, included in the study that delivered significantly better results than the baseline. The rest of the augmented models, homepage and external bookmark CBF models and all three external-source-augmented CBCF models, did not yield any significant improvement on recommendation.

In study 2, cold-start context, homepage, bibliography, and external bookmark CBF models performed significantly better than baselines in the beginning-stage cold-start situation where there was no bookmarked talk in the user profile. After that, only homepage and bibliography CBF models significantly outperformed the CBF baselines to the latter stages (0 – 2 bookmarked talks for the homepage CBF model and 0 – 16 ones for the bibliography one). The CBCF models with external-source-augmentation significantly boosted the recommendation performance only in the beginning stage where there was no bookmarked talk in the user profile.

In the recommendation fusion CN3 study in chapter eight, most fusion models performed statistically better than CBF baselines. However, only nine of them significantly outperformed CBCF baselines. Examination of the fusion methods revealed, CombMNZ models produced more reliable results than CombSUM counterparts because the CombMNZ method relied on the number of sources. Because there were only two recommendation sources, the target talks were more likely to be ranked higher if they received high recommendation scores and confirmations from both sources.

The external validity study also confirmed previous CN3 findings from chapter six to chapter eight. The validity of study 1 showed that the bibliography-augmented CBF model was the only one that statistically outperformed the baseline CBF counterpart.

In conclusion, the bibliography source was the only one that improved recommendation for the content-based recommender type. The bibliography source seemed to contain more relevant and less irrelevant unigram terms regarding user interests than the other two sources. As a result, injecting bibliography terms into the user profile provided improvement on content-based recommendation.

Models with homepage augmentation were expected to perform the worst. Surprisingly, they performed slightly higher than baselines in the global-impact study, and they boosted performance in the cold-start situations. One observation of this source was that the homepage contained diverse content segmentations such as user publications. As a result, injecting this source into the user profile was more likely to produce good results.

However, CBF models with external bookmark augmentation performed poorly. They were expected to improve performance of content-based recommendation. One observation was they contained more extra terms outside of the CN3 corpus than the other two sources. Those

terms not only introduced noises into the user profile but also diluted the values of relevant terms in the user profile as explained in the discussion of chapter six.

On the CBCF side, there was no source capable of increasing the performance of recommendations beyond the baseline. One explanation was the user profile consisted of enough relevant unigram terms in searching for nearest same-tasted peers. With extra information augmented, the user profile did not gain extra ability to find better peers. Also, the difference between results of CBF models and CBCF models were prominent. Results of CBCF models did not gain impact from the strong assumption of experimental design as opposed to CBF models. The reasons were deduced from user behaviors. Nearest peers, who had similar interests to the particular user, tended to attend the same conferences. As a result, CBCF models were robust to the experimental design effect. From another perspective, CBCF models were regarded as an artifact of this study.

12.2.2 CoMeT DATASET

The results of the user study of the CoMeT system were not exactly the same as ones produced by the cross validation study of the CN3 dataset. It was speculated that differences were caused by the difference in data collection of two systems and two different assumption of their ground truth. The CN3 dataset was sparse. Most of the CN3 users bookmarked talks from conferences they attended. In the CN3 experimental design, the remaining talks left in the corpus were assumed to not interest users. Because users did not attend these particular conferences, their ground truth was unknown. It is a possibility that users might have bookmarked some of the non-bookmarked talks had they attended those conferences. On the other hand, in the CoMeT user

study, participants had to provide feedback on every single talk in both training and test sets whether or not they were bookmarked. Therefore, the CoMeT dataset was more complete than the CN3 dataset, and there was an ample amount of bookmarked talks for constructing good baseline models in the CoMeT study. Consequently, the MAP results of baseline models in all three CoMeT sub-studies were already high. Thus, adding extra information into user profiles did not help improve the performance of the models. Secondly, the nature of the two systems is different. CN3 hosts talks of research conferences. The CN3 talks tend to be more coherent and related. While CoMeT talks are more diversified. The interest areas of CN3 talks range from computer sciences, medical and health sciences, to liberal arts and philosophy. Another observation was the bibliography source helped improve the performance in the CN3 studies, but it did not in the CoMeT case. Most CN3 users were scholars or senior researchers. They had already published quite a few papers (median: 27). On the other hand, CoMeT users were PhD students or junior researchers (median: 9). Coupled with the nature of the CN3 system, it was more likely that CN3 users were more experienced and tended to attend talks related to their research interests. On the other hand, CoMeT tended to attend talks in order to explore beyond their researches.

12.2.3 Cross-System SciNet-CoMeT DATASET

The cross-system studies of the SciNet-CoMeT dataset were performed as retrospective studies to find new external user sources that were able to improve recommendation performance. In the global-impact of user model study, there were no significantly different results from any transfer user model and the baseline. The results showed that the user profiles, with enough information

about user interest, did not receive any good impact from augmenting them with either explicitly curated user models or traditional implicit un-processed user models. However, the transfer of explicitly curated user models yielded better results than alternative transfer approaches and the baseline. The reason for the good performance of open curated models can be analyzed by examining similarities between the different types of transferred information and the information contained in the bookmarked and non-bookmarked CoMeT talks. Consequently, the result of the cold-start study suggested that fewer but relevant user terms or keywords were the key factor to improve the performance of recommendation. Subsequent study showed machine-learning approaches, such as feature selection, were able to generate better a representation of user interests.

13.0 SUMMARY AND CONCLUSIONS

This chapter starts with a summary of the results and the implications of this research and concludes with a discussion of future work.

13.1 SUMMARY OF RESULTS

Studies were conducted to explore whether external sources were helpful to improve research talk recommendations in small research communities. Research talk recommendations face two main challenges. One is the “under-contribution” situation. In this situation, the social system receives a few number of user contributions such as bookmarks or ratings. The second challenge is short life time-span of research talks. Research talks have little or no recommendation value after they take place. Under both conditions, user contribution to the social systems, which is sparse in the “under-contribution” situation, is even sparser when utilizing these contributions to generate recommendations. As a result, homepage, bibliography, and bookmarked scientific papers were selected as external sources for this dissertation. They were used to generate external-source-augmented recommendations. Moreover, an extra source, SciNet user profiles and search logs, was used to create cross-system comparisons. There were three aspects this research investigated. First aspect was the external-source-augmented recommendation study.

Second was the cold-start problem in the recommendation. The last one was the recommendation fusion. There were also four kinds of investigations: k-fold cross-validation for CN3, external validity of CN3 study, user study for CoMeT system, and cross-system CoMeT-SciNet cross validation.

In the external-source-augmented recommendation, the five-fold cross-validation in the CN3 study showed only one model, the full-text 5-NN.PO CN3-term model with bibliography augmentation, outperformed the content-based baseline model. The rest of the experimental models did not perform better than the baselines in either content-based or collaborative filtering recommender types.

In the cold-start problem study, the five-fold cross-validation CN3 research showed that all three models with external source augmentation were able to help alleviate the cold-start problem when users did not have any bookmarked talk in either content-based or content-boosted collaborative filtering recommender types. After that, the homepage-augmented content-based model was able to outperform the content-based baseline model until users had two bookmark talks in their profile. The bibliography-augmented content-based model performed very well. It was able to perform statistically better than the content-based model from no bookmark at all to sixteen bookmarks.

In the recommendation fusion study, the five-fold cross-validation CN3 study showed that fusion of external-source-augmented models improved the performance of recommendations. Almost of all confusion models outperformed the content-based baselines. Also, as never before seen in previous studies, nine out of all fusion models were able to perform statistically better than the content-boosted collaborative filtering baselines.

The external validity study confirmed the results from three CN3 cross-validation studies. In the reevaluation of Study 1, as expected, the bibliography-augmented CBF model significantly improved the performance. In the re-evaluation of the cold-start problem, homepage, bibliography, and external-bookmark augmentation models helped significantly improve recommendation results when users had no bookmarked talk. In the re-evaluation of the recommendation fusion study, the recommendation fusion results of the external validity confirmed the ones from Study 3.

However, in the CoMeT user study, the research showed there was no model that outperformed baselines, similar to the CN3 study. In this study, the research revealed a mix message. In attending conference situations, user publications played a role in improving content-based recommendation. Users seemed to select talks that related to their bibliography. However, in the university or research environment, external source did not help improve talk recommendations.

In the cold-start problem study, like the CN3 study, the experimental models outperformed baselines when there was no bookmark in their profile. There was no experimental model that outperformed baselines when users started bookmarking talks. As a result, the external sources provided the ability to alleviate the cold-start effect on talk recommendations. Moreover, they were more helpful in recommending talks in conference venues than talks in academia in the cold-start problem.

In the recommendation fusion in the CoMeT user study, even though the fusion models improved the performance of non-fusion models, they did not perform statistically better than either baseline in all three measures, including MAP, nDCG on Relevance, and nDCG on Novelty.

In the cross-system user study, we explored whether the user model transfer from SciNet system could help improve recommendations in the CoMeT system. In this study, the transfer models did not improve recommendation in general situations. However, the transfer models, based on open user models, showed positive impacts on the performance of recommendations in the cold-start situation. They significantly outperformed the baseline in the cold-start situation.

In summary, some of the external source augmentations provided positive value to talk recommendations in small research communities. However, not all external sources were useful. For example, the research showed adding external scholarly bookmarked papers into recommending models in both cross-validation and user studies did not improve the recommendation performance but harmed it. Also, augmenting user models with external sources improved talk recommendations in the early stages of cold-start situations. Combining those external-source-augmented recommendations with right weights increased the performance of mean average precision in the cross-validation but made no difference in the CoMeT user study.

In the transfer user model study, adding full-text of bookmarked articles or displayed articles from SciNet and CoMeT did not improve performances of the three measures but harmed them. While adding all unigrams from bookmarked or displayed papers into the user model harms the performance, adding interacting or displayed keywords from open user models helped improve recommendation in the early stages of cross-model transition situations. One lesson learned from the results was that user information was better processed; for example interacting or displayed keywords from the SciNet open user model, seemed to be good external sources to help improve recommendations.

13.2 IMPLICATIONS AND LESSONS LEARNED

13.2.1 Impact of External Source Augmentation on Recommendation

Conducted studies provided evidence that external source might be useful when recommending research talks. Among all of them, bibliography can deliver a largest improvement to CBF recommendations if users are experienced scholars or researchers and have considerable list of publications. When the number of bookmarked talks in the user profile is large, augmenting CBF models does not provide any advantage. However, in situations where there are no or few bookmarked talks, the external sources help CBF models smooth the severity of the cold-start problem.

The majority of external sources explored in this study are considered implicit, un-curated user information. They might not be reliable representations of user interests. We learned that transferring explicit user model is a good alternative to transferring raw information. It boosts the CBF recommendation, especially in the cold-start context.

As we learned, CBCF models yield relatively high results without adding external sources. However, performance of CBCF depends on peers to recommend items. When a number of peers sharing similar interests are too small like they were in CoMeT-SciNet, CBCF might deliver unreliable recommendation or no recommendation at all.

We learned that when recommendations are good, fusing them gives better results. The method of fusion is also a factor to be considered. In our case, CombMNZ has an edge over CombSUM.

13.2.2 The Effect of Experimental Design Assumption

In the CN3 study, strong assumptions on user ground truth were made. It was assumed that talk users without bookmarks were all non-interested. The problem is, however, that users did not have a chance to provide feedback on talks from conferences they had not attended. This apparently contaminated the test set with some talks cannot serve as a reliable ground truth. In future studies, the experimental design performed on a partial-ground-truth dataset, like CN3, should incorporate only items that users had a chance to see and bookmark in the test set.

13.3 FUTURE WORK

This research showed that external sources benefit talk recommendations in small research communities. The research showed that well-processed external user information or explicit open user models work better than raw data or implicit user models. As a result, research challenges will be pursued to exploit the use of external sources to improve talk recommendations. These include:

- Feature Selection
- Textual Information Analysis and Topic Classification
- Incorporating External Sources into Model-based Recommendation Techniques

Advanced machine learning techniques can be introduced into future studies as a way to improve the performance of recommending models. One simple approach is features selection. According to results from transfer user model experiments conducted in chapter eleven, the study

showed that transferred user models provided positive impact to recommendations in the cold-start setting. The user information from transferred user models contained only a few hundred terms in the manipulated keyword logs and a few thousand terms in the shown keyword logs, while raw data from bookmarked or displayed articles consisted of around thirty thousand terms. Clearly, there is a lot of noise in the raw data from SciNet bookmarked or displayed articles. The forward selection or backward elimination or combination of the two can be used as a way to search for relevant terms in the external sources information.

External information sources contain many different types of content. For example, personal webpages include information about many user aspects, such as personal information, teaching courses, publications, favorite hobbies, and other characteristics. Therefore, it could be valuable to separate them into groups or sections of content and recommend talks based on each of these groups. Prolific scholars publish a considerable number of scientific articles. There is a possibility that their publications belong to many different topics or domains. Therefore, conducting content analysis on those bibliographies is a challenge to examine the extent those cluster or topic classifications are able to improve the recommendations.

Unlike the simple approach, like feature selection, the advanced, state-of-the-art, graphical models, such as Latent Dirichlet Allocation (LDA) models, can be an inference method that incorporates the external sources as extra parameters to infer user interests and generate recommendations. Another interesting model is the Multitask Learning technique. The Multitask Learning technique can be considered as a fusion method of multiple external sources. It considers each external-source-augmented model as a task to recommend talk and assumes each task has incomplete data. The approach exploits the fact that these tasks are related and recommends most preferred talks.

BIBLIOGRAPHY

Abel, F., Herder, E., Houben, G. J., Henze, N., & Krause, D. (2013). Cross-system user modeling and personalization on the social web. *User Modeling and User-Adapted Interaction*, *23*(2-3), 169-209.

Adomavicius, G., & Tuzhilin, A. (2005). Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *Knowledge and Data Engineering, IEEE Transactions on*, *17*(6), 734-749.

Agichtein, E., Brill, E., & Dumais, S. (2006, August). Improving web search ranking by incorporating user behavior information. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 19-26). ACM.

Bailey, T. L., & Elkan, C. (1994). Fitting a mixture model by expectation maximization to discover motifs in bipolymers.

Baglama, J., & Reichel, L. (2014). irlba: Fast Partial SVD by Implicitly-Restarted Lanczos Bidiagonalization. R Package Version 1.0. 3.

Basu, C., Hirsh, H., & Cohen, W. (1998, July). Recommendation as classification: Using social and content-based information in recommendation. In *Proceedings of the national conference on artificial intelligence* (pp. 714-720). John Wiley & Sons LTD.

Baeza-Yates, R., & Ribeiro-Neto, B. (1999). *Modern information retrieval* (Vol. 463). New York: ACM press.

Belkin, N. J., & Croft, W. B. (1992). Information filtering and information retrieval: two sides of the same coin?. *Communications of the ACM*, *35*(12), 29-38.

Belkin, N. J., Kantor, P., Fox, E. A., & Shaw, J. A. (1995). Combining the evidence of multiple query representations for information retrieval. *Information Processing & Management*, *31*(3), 431-448.

Berkovsky, S., Kuflik, T., & Ricci, F. (2008). Mediation of user models for enhanced personalization in recommender systems. *User Modeling and User-Adapted Interaction*, *18*(3), 245-286.

Berkovsky, S., Kuflik, T., & Ricci, F. (2009). Cross-representation mediation of user models. *User Modeling and User-Adapted Interaction*, *19*(1-2), 35-63.

Billsus, D., & Pazzani, M. J. (1998, July). Learning collaborative information filters. In *Proceedings of the fifteenth international conference on machine learning* (Vol. 54, p. 48).

Billsus, D., & Pazzani, M. J. (2000). User modeling for adaptive news access.*User modeling and user-adapted interaction*, *10*(2-3), 147-180.

Bloedorn, E., Mani, I., & MacMillan, T. R. (1996, March). Representational issues in machine learning of user profiles. In *Proceedings of the Thirteenth National Conference on Artificial Intelligence and Eighth Innovative Applications of Artificial Intelligence Conference*.

Bogers, A. M. (2009). *Recommender systems for social bookmarking*. Diss. University of Tilburg, 2009.

Bollacker, K. D., Lawrence, S., & Giles, C. L. (1999, August). A system for automatic personalized tracking of scientific literature on the web. In*Proceedings of the fourth ACM conference on Digital libraries* (pp. 105-113). ACM.

Bollacker, K. D., Lawrence, S., & Giles, C. L. (2000). Discovering relevant scientific literature on the web. *Intelligent Systems and their Applications, IEEE*, *15*(2), 42-47.

Breese, J. S., Heckerman, D., & Kadie, C. (1998, July). Empirical analysis of predictive algorithms for collaborative filtering. In *Proceedings of the Fourteenth conference on Uncertainty in artificial intelligence* (pp. 43-52). Morgan Kaufmann Publishers Inc..

Breiman, L. (1996). Bagging predictors. *Machine learning*, *24*(2), 123-140.

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5-32.

Brusilovsky, P., Parra, D., Sahebi, S., & Wongchokprasitti, C. (2010, October). Collaborative information finding in smaller communities: The case of research talks. In *Collaborative Computing: Networking, Applications and Worksharing (CollaborateCom), 2010 6th International Conference on* (pp. 1-10). IEEE.

Burke, R. (1999, July). Integrating knowledge-based and collaborative-filtering recommender systems. In *Proceedings of the Workshop on AI and Electronic Commerce* (pp. 69-72).

Burke, R. (2002). Hybrid recommender systems: Survey and experiments. *User modeling and user-adapted interaction*, *12*(4), 331-370.

Carmagnola, F., Cena, F., & Gena, C. (2011). User model interoperability: a survey. *User Modeling and User-Adapted Interaction*, *21*(3), 285-331.

Chen, P. M., & Kuo, F. C. (2000). An information retrieval system based on a user profile. *Journal of Systems and Software*, *54*(1), 3-8.

Chen, L., & Sycara, K. (1998, May). WebMate: a personal agent for browsing and searching. In *Proceedings of the second international conference on Autonomous agents* (pp. 132-139). ACM.

Chiu, B. C., & Webb, G. I. (1998). Using decision trees for agent modeling: improving prediction performance. *User Modeling and User-Adapted Interaction,8*(1-2), 131-152.

Choochaiwattana, Worasit. *Using social annotations to improve web search*. Diss. University of Pittsburgh, 2008.

Chung, R., Sundaram, D., & Srinivasan, A. (2007, August). Integrated personal recommender systems. In *Proceedings of the ninth international conference on Electronic commerce* (pp. 65-74). ACM.

Cremonesi, P., Tripodi, A., & Turrin, R. (2011, December). Cross-domain recommender systems. In *Data Mining Workshops (ICDMW), 2011 IEEE 11th International Conference on* (pp. 496-503). IEEE.

Croft, W. B. (2002). *Combining approaches to information retrieval* (pp. 1-36). Springer US.

Croft, W. B., & Thompson, R. H. (1987). I3R: A new approach to the design of document retrieval systems. *Journal of the american society for information science*, *38*(6), 389-404.

Dasan, V. S. (1998). *U.S. Patent No. 5,761,662*. Washington, DC: U.S. Patent and Trademark Office.

Daoud, M., Tamine, L., & Boughanem, M. (2010). A personalized graph-based document ranking model using a semantic user profile. In *User Modeling, Adaptation, and Personalization* (pp. 171-182). Springer Berlin Heidelberg.

Deerwester, S. C., Dumais, S. T., Landauer, T. K., Furnas, G. W., & Harshman, R. A. (1990). Indexing by latent semantic analysis. *JASIS*, *41*(6), 391-407.

Diederich, J., & Iofciu, T. (2006, October). Finding communities of practice from user profiles based on folksonomies. In *Proceedings of the 1st International Workshop on Building Technology Enhanced Learning solutions for Communities of Practice* (Vol. 213).

Falagas, M. E., Pitsouni, E. I., Malietzis, G. A., & Pappas, G. (2008). Comparison of PubMed, Scopus, web of science, and Google scholar: strengths and weaknesses. *The FASEB Journal*, 22(2), 338-342.

Farzan, R., & Brusilovsky, P. (2008). AnnotatEd: A social navigation and annotation service for web-based educational resources. *New Review of Hypermedia and Multimedia*, *14*(1), 3-32.

Farzan, R., & Brusilovsky, P. (2005). Social navigation support through annotation-based group modeling. In *User Modeling 2005* (pp. 463-472). Springer Berlin Heidelberg.

Farzan, R., DiMicco, J. M., Millen, D. R., Dugan, C., Geyer, W., & Brownholtz, E. A. (2008, April). Results from deploying a participation incentive mechanism within the enterprise. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 563-572). ACM.

Fernández-Tobías, I., Cantador, I., Kaminskas, M., & Ricci, F. (2012, June). Cross-domain recommender systems: A survey of the state of the art. In*Spanish Conference on Information Retrieval*.

Fortunato, S. (2010). Community detection in graphs. *Physics Reports*, *486*(3), 75-174.

Fossati, M., Giuliano, C., & Tummarello, G. (2012, October). Semantic Network-driven News Recommender Systems: a Celebrity Gossip Use Case. In *SeRSy* (pp. 25-36).

Freund, Y., & Schapire, R. E. (1996, July). Experiments with a new boosting algorithm. In *MACHINE LEARNING-INTERNATIONAL WORKSHOP THEN CONFERENCE*- (pp. 148-156). MORGAN KAUFMANN PUBLISHERS, INC..

Gan, G., Ma, C., & Wu, J. (2007). *Data clustering: theory, algorithms, and applications* (Vol. 20). Society for Industrial and Applied Mathematics.

Gauch, S., Speretta, M., Chandramouli, A., & Micarelli, A. (2007). User profiles for personalized information access. In *The adaptive web* (pp. 54-89). Springer Berlin Heidelberg.

Gentili, G., Micarelli, A., & Sciarrone, F. (2003). Infoweb: An adaptive information filtering system for the cultural heritage domain. *Applied Artificial Intelligence*, *17*(8-9), 715-744.

Godoy, D., Schiaffino, S., & Amandi, A. (2004). Interface agents personalizing Web-based tasks. *Cognitive Systems Research*, *5*(3), 207-222.

Goldberg, D., Nichols, D., Oki, B. M., & Terry, D. (1992). Using collaborative filtering to weave an information tapestry. *Communications of the ACM*, *35*(12), 61-70.

Golub, G. H., & Reinsch, C. (1970). Singular value decomposition and least squares solutions. *Numerische Mathematik*, *14*(5), 403-420.

Gori, M., & Pucci, A. (2006, December). Research paper recommender systems: A random-walk based approach. In *Web Intelligence, 2006. WI 2006. IEEE/WIC/ACM International Conference on* (pp. 778-781). IEEE.

Handcock, M. S., Raftery, A. E., & Tantrum, J. M. (2007). Model-based clustering for social networks. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *170*(2), 301-354.

Hartigan, J. A., & Wong, M. A. (1979). Algorithm AS 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, *28*(1), 100-108.

Heckmann, D., Schwartz, T., Brandherm, B., Schmitz, M., & von Wilamowitz-Moellendorff, M. (2005). Gumo–the general user model ontology. In *User modeling 2005* (pp. 428-432). Springer Berlin Heidelberg.

Herlocker, J. L., Konstan, J. A., Terveen, L. G., & Riedl, J. T. (2004). Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems (TOIS)*, *22*(1), 5-53.

Hofmann, T. (2003, July). Collaborative filtering via gaussian probabilistic latent semantic analysis. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval* (pp. 259-266). ACM.

Ingwersen, P., & Järvelin, K. (2005). *The turn: Integration of information seeking and retrieval in context* (Vol. 18). Springer.

Jain, A. K., & Dubes, R. C. (1988). *Algorithms for clustering data*. Prentice-Hall, Inc..

Jarvelin, K., and Kekalainen., J. *IR evaluation methods for retrieving highly relevant documents.* In Proceedings of the 23rd annual international ACM SIGIR conference on Research and Development on Information Retrieval, July 24-28, 2000, Athens, Greece.

Joachims, T. (2006, August). Training linear SVMs in linear time. In*Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 217-226). ACM.

Kettenring, J. R. (2006). The practice of cluster analysis. *Journal of classification*, *23*(1), 3-30.

Katz, G., Ofek, N., Shapira, B., Rokach, L., & Shani, G. (2011, October). Using Wikipedia to boost collaborative filtering techniques. In *Proceedings of the fifth ACM conference on Recommender systems* (pp. 285-288). ACM.

Kiang, M. Y. (2001). Extending the Kohonen self-organizing map networks for clustering analysis. *Computational Statistics & Data Analysis*, *38*(2), 161-180.

Klema, V., & Laub, A. (1980). The singular value decomposition: Its computation and some applications. *Automatic Control, IEEE Transactions on*,*25*(2), 164-176.

Kohonen, T. (1990). The self-organizing map. *Proceedings of the IEEE*, *78*(9), 1464-1480.

Koren, Y. (2008, August). Factorization meets the neighborhood: a multifaceted collaborative filtering model. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 426-434). ACM.

Koren, Y., Bell, R., & Volinsky, C. (2009). Matrix factorization techniques for recommender systems. *Computer*, *42*(8), 30-37.

Kretschmer, H. (1994). Coauthorship networks of invisible colleges and institutionalized communities. *Scientometrics*, 30(1), 363-369.

Lee, D. H. *Recommendations based on users' various social networks.* Diss. University of Pittsburgh, 2013.

Lee, J. H. (1997, December). Analyses of multiple evidence combination. In *ACM SIGIR Forum* (Vol. 31, No. SI, pp. 267-276). ACM.

Lee Rodgers, J., & Nicewander, W. A. (1988). Thirteen ways to look at the correlation coefficient. *The American Statistician*, *42*(1), 59-66.

Lemire, D., & Maclachlan, A. (2005). Slope one predictors for online rating-based collaborative filtering. *Society for Industrial Mathematics*, *5*, 471-480.

Li, B., Yang, Q., & Xue, X. (2009, July). Can Movies and Books Collaborate? Cross-Domain Collaborative Filtering for Sparsity Reduction. In *IJCAI* (Vol. 9, pp. 2052-2057).

Liu, Y. T., Liu, T. Y., Qin, T., Ma, Z. M., & Li, H. (2007, May). Supervised rank aggregation. In *Proceedings of the 16th international conference on World Wide Web* (pp. 481-490). ACM.

Liu, F., Yu, C., & Meng, W. (2002, November). Personalized web search by mapping user queries to categories. In *Proceedings of the eleventh international conference on Information and knowledge management* (pp. 558-565). ACM.

Loizou, A. (2009). *How to recommend music to film buffs: enabling the provision of recommendations from multiple domains* (Doctoral dissertation, University of Southampton).

MacQueen, J. (1967, June). Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability* (Vol. 1, No. 281-297, p. 14).

Maltz, D., & Ehrlich, K. (1995, May). Pointing the way: active collaborative filtering. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (pp. 202-209). ACM Press/Addison-Wesley Publishing Co..

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval* (Vol. 1). Cambridge: Cambridge University Press.

Matthijs, N., & Radlinski, F. (2011, February). Personalizing web search using long term browsing history. In *Proceedings of the fourth ACM international conference on Web search and data mining* (pp. 25-34). ACM.

McNee, S. M., Albert, I., Cosley, D., Gopalkrishnan, P., Lam, S. K., Rashid, A. M., ... & Riedl, J. (2002, November). On the recommending of citations for research papers. In *Proceedings of the 2002 ACM conference on Computer supported cooperative work* (pp. 116-125). ACM.

McNee, S. M., Kapoor, N., & Konstan, J. A. (2006, November). Don't look stupid: avoiding pitfalls when recommending research papers. In *Proceedings of the 2006 20th anniversary conference on Computer supported cooperative work* (pp. 171-180). ACM.

Mehta, B. (2008). *Cross System Personalization: Enabling personalization across multiple systems* (Doctoral dissertation, Universität Duisburg-Essen, Fakultät für Ingenieurwissenschaften» Ingenieurwissenschaften-Campus Duisburg» Abteilung Informatik und Angewandte Kognitionswissenschaft» Informationssysteme).

Melville, P., Mooney, R. J., & Nagarajan, R. (2002, July). Content-boosted collaborative filtering for improved recommendations. In *Proceedings of the National Conference on Artificial Intelligence* (pp. 187-192). Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999.

Middleton, S. E., Alani, H., & De Roure, D. C. (2002). Exploiting synergy between ontologies and recommender systems. *arXiv preprint cs/0204012*.

Middleton, S. E., De Roure, D. C., & Shadbolt, N. R. (2001, October). Capturing knowledge of user preferences: ontologies in recommender systems. In *Proceedings of the 1st international conference on Knowledge capture* (pp. 100-107). ACM.

Middleton, S. E., Shadbolt, N. R., & De Roure, D. C. (2004). Ontological user profiling in recommender systems. *ACM Transactions on Information Systems (TOIS)*, *22*(1), 54-88.

Minkov, E., Charrow, B., Ledlie, J., Teller, S., & Jaakkola, T. (2010, October). Collaborative future event recommendation. In *Proceedings of the 19th ACM international conference on Information and knowledge management* (pp. 819-828). ACM.

Mnih, A., & Salakhutdinov, R. (2007). Probabilistic matrix factorization. In*Advances in neural information processing systems* (pp. 1257-1264).

Montague, M., & Aslam, J. A. (2001, October). Relevance score normalization for metasearch. In *Proceedings of the tenth international conference on Information and knowledge management* (pp. 427-433). ACM.

Moon, T. K. (1996). The expectation-maximization algorithm. *Signal Processing Magazine, IEEE*, *13*(6), 47-60.

Newman, M. E. (2001). The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences*, 98(2), 404-409.

Newman, M. E. (2004). Fast algorithm for detecting community structure in networks. *Physical review E*, 69(6), 066133.

Page, L., Brin, S., Motwani, R., & Winograd, T. (1999). The PageRank citation ranking: bringing order to the web.

Palla, G., Derényi, I., Farkas, I., & Vicsek, T. (2005). Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, *435*(7043), 814-818.

Pan, W., Liu, N. N., Xiang, E. W., & Yang, Q. (2011, July). Transfer learning to predict missing ratings via heterogeneous user feedbacks. In *IJCAI Proceedings-International Joint Conference on Artificial Intelligence* (Vol. 22, No. 3, p. 2318).

Parra, D., & Sahebi, S. (2013). Recommender Systems: Sources of Knowledge and Evaluation Metrics. In *Advanced Techniques in Web Intelligence-2* (pp. 149-175). Springer Berlin Heidelberg.

Pavlov, D., & Pennock, D. M. (2002, December). A maximum entropy approach to collaborative filtering in dynamic, sparse, high-dimensional domains. In *Neural Information Processing Systems* (pp. 1441-1448).

Pazzani, M. J. (1999). A framework for collaborative, content-based and demographic filtering. *Artificial Intelligence Review*, *13*(5-6), 393-408.

Pelleg, D., & Moore, A. (2000, June). X-means: Extending k-means with efficient estimation of the number of clusters. In *Proceedings of the seventeenth international conference on machine learning* (Vol. 1, pp. 727-734).

Radicchi, F., Castellano, C., Cecconi, F., Loreto, V., & Parisi, D. (2004). Defining and identifying communities in networks. *Proceedings of the National Academy of Sciences of the United States of America, 101*(9), 2658-2663.

Rafter, R., & Smyth, B. (2001, August). Passive profiling from server logs in an online recruitment environment. In *Proceedings of IJCAI Workshop on Intelligent Techniques for Web Personalization (ITWP2001), Seattle, Washington, USA.*

Renda, M. E., & Straccia, U. (2003, March). Web metasearch: rank vs. score based rank aggregation methods. In *Proceedings of the 2003 ACM symposium on Applied computing* (pp. 841-846). ACM.

Rennie, J. D., & Srebro, N. (2005, August). Fast maximum margin matrix factorization for collaborative prediction. In *Proceedings of the 22nd international conference on Machine learning* (pp. 713-719). ACM.

Rodriguez, J. J., Kuncheva, L. I., & Alonso, C. J. (2006). Rotation forest: A new classifier ensemble method. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, *28*(10), 1619-1630.

Resnick, P., Iacovou, N., Suchak, M., Bergstrom, P., & Riedl, J. (1994, October). GroupLens: an open architecture for collaborative filtering of netnews. In *Proceedings of the 1994 ACM conference on Computer supported cooperative work* (pp. 175-186). ACM.

Ruotsalo, T., Peltonen, J., Eugster, M., Głowacka, D., Konyushkova, K., Athukorala, K., ... & Kaski, S. (2013, October). Directing exploratory search with interactive intent modeling. In *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management* (pp. 1759-1764). ACM.

Sahebi, S. & Brusilovsky, P. (2013). Cross-Domain Recommendation in a Cold-Start Context: The impact of User Profile Size on the Quality of Recommendation. In *Proceedings of The 21st Conference on User Modeling, Adaptation and Personalization.* Springer Berlin Heidelberg.

Sahebi, S., Wongchokprasitti, C., & Brusilovsky, P. (2010, September). Recommending research colloquia: a study of several sources for user profiling. In *Proceedings of the 1st International Workshop on Information Heterogeneity and Fusion in Recommender Systems* (pp. 32-38). ACM.

Salton, G., & McGill, M. J. (1983). Introduction to modern information retrieval.

Sarwar, B., Karypis, G., Konstan, J., & Riedl, J. (2001, April). Item-based collaborative filtering recommendation algorithms. In *Proceedings of the 10th international conference on World Wide Web* (pp. 285-295). ACM.

Schafer, J. B., Frankowski, D., Herlocker, J., & Sen, S. (2007). Collaborative filtering recommender systems. In *The adaptive web* (pp. 291-324). Springer Berlin Heidelberg.

Schaffer, J. D., Lee, K. P., Kurapati, K., & Gutta, S. (2010). *U.S. Patent No. 7,721,310*. Washington, DC: U.S. Patent and Trademark Office.

Schein, A. I., Popescul, A., Ungar, L. H., & Pennock, D. M. (2002, August). Methods and metrics for cold-start recommendations. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 253-260). ACM.

Schwab, I., Pohl, W., & Koychev, I. (2000, January). Learning to recommend from positive evidence. In *Proceedings of the 5th international conference on Intelligent user interfaces* (pp. 241-247). ACM.

Shardanand, U., & Maes, P. (1995, May). Social information filtering: algorithms for automating "word of mouth". In *Proceedings of the SIGCHI conference on Human factors in computing systems* (pp. 210-217). ACM Press/Addison-Wesley Publishing Co..

Shi, Y., Larson, M., & Hanjalic, A. (2011). Tags as bridges between domains: Improving recommendation with tag-induced cross-domain collaborative filtering. In *User Modeling, Adaption and Personalization* (pp. 305-316). Springer Berlin Heidelberg.

Semeraro, G., Lops, P., Basile, P., & de Gemmis, M. (2009, October). Knowledge infusion into content-based recommender systems. In *Proceedings of the third ACM conference on Recommender systems* (pp. 301-304). ACM.

Sen, P. (2006). Complexities of social networks: A Physicist's perspective. *arXiv preprint physics/0605072*.

Smith, K. A., & Ng, A. (2003). Web page clustering using a self-organizing map of user navigation patterns. *Decision Support Systems*, *35*(2), 245-256.

Sosnovsky, S., Brusilovsky, P., Yudelson, M., Mitrovic, A., Mathews, M., & Kumar, A. (2009). *Semantic integration of adaptive educational systems* (pp. 134-158). Springer Berlin Heidelberg.

Sugiyama, K., Hatano, K., & Yoshikawa, M. (2004, May). Adaptive web search based on user profile constructed without any effort from users. In *Proceedings of the 13th international conference on World Wide Web* (pp. 675-684). ACM.

Thorndike, R. L. (1953). Who belongs in the family?. *Psychometrika*, *18*(4), 267-276.

Vargas, S., & Castells, P. (2011, October). Rank and relevance in novelty and diversity metrics for recommender systems. In *Proceedings of the fifth ACM conference on Recommender systems* (pp. 109-116). ACM.

Van Loan, C. F. (1976). Generalizing the singular value decomposition. *SIAM Journal on Numerical Analysis*, *13*(1), 76-83.

Velden, T., Haque, A. U., & Lagoze, C. (2010). A new approach to analyzing patterns of collaboration in co-authorship networks: mesoscopic analysis and interpretation. *Scientometrics*, 85(1), 219-242.

Vesanto, J., & Alhoniemi, E. (2000). Clustering of the self-organizing map.*Neural Networks, IEEE Transactions on*, *11*(3), 586-600.

Wall, M., Rechtsteiner, A., & Rocha, L. (2003). Singular value decomposition and principal component analysis. *A practical approach to microarray data analysis*, 91-109.

Wang, C., & Blei, D. M. (2011, August). Collaborative topic modeling for recommending scientific articles. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 448-456). ACM.

White, R. W., Ruthven, I., & Jose, J. M. (2002, August). Finding relevant documents using top ranking sentences: an evaluation of two alternative schemes. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 57-64). ACM.

Wongchokprasitti, C., & Brusilovsky, P. (2007, June). Newsme: A case study for adaptive news systems with open user model. In *Autonomic and Autonomous Systems, 2007. ICAS07. Third International Conference on* (pp. 69-69). IEEE.

Wongchokprasitti, C., Brusilovsky, P., & Parra-Santander, D. (2010). *Conference Navigator 2.0: community-based recommendation for academic conferences*. In: Workshop on Social Reminder Systems (SRS '10), 7 February 2010, Hong Kong, China.

Wongchokprasitti, C., Peltonen, J., Ruotsalo, T., Bandyopadhyay, P., Jacucci, G., & Brusilovsky, P. (2015). User Model in a Box: Cross-System User Model Transfer for Resolving Cold Start Problems. In *User Modeling, Adaptation and Personalization* (pp. 289-301). Springer International Publishing.

Xu, R., & Wunsch, D. (2005). Survey of clustering algorithms. *Neural Networks, IEEE Transactions on*, *16*(3), 645-678.

Xue, G. R., Han, J., Yu, Y., & Yang, Q. (2009). User language model for collaborative personalized search. *ACM Transactions on Information Systems (TOIS)*, *27*(2), 11.

Zhou, T., Ren, J., Medo, M., & Zhang, Y. C. (2007). Bipartite network projection and personal recommendation. *Physical Review E*, *76*(4), 046115.

Zigoris, P., & Zhang, Y. (2006, November). Bayesian adaptive user profiling with explicit & implicit feedback. In *Proceedings of the 15th ACM international conference on Information and knowledge management* (pp. 397-404). ACM.