Utilizing Social Breadcrumbs for User Profiling in Personalization

Arjumand Younus

College of Engineering and Informatics
National University of Ireland, Galway

Department of Informatics, Systems and Communications
University of Milano-Bicocca, Milan

A thesis submitted for the degree of
Doctor of Philosophy

2015

Acknowledgements

I wish to express immense amount of gratitude to my supervisors Dr. Colm O'Riordan and Dr. Gabriella Pasi. Their words of motivation and encouragement helped in uplifting my spirit and motivated me strongly for research. Moreover, their useful advice and feedback on various aspects of my PhD research helped me polish it to the best of my abilities. In particular, I would like to express my special gratitude to Dr. Colm O'Riordan for his constant endeavours throughout the journey of this PhD which is something I will forever cherish, and look back as one of the fondest memories of my life. Indeed, one is very lucky to find a good friend such as Dr. Colm O'Riordan in a supervisor.

I am very grateful to the Hardiman scholarship committee within National University of Ireland, Galway who considered me worthy of such a prestigious scholarship. It surely would not have been possible to complete the research conducted in this thesis without their support. Their constant endeavours at pursuing a follow-up of how I was getting along in my research kept me on my feet, and I am indebted to them for such continuous motivation.

I am grateful to Josephine Griffith of National University of Ireland, Galway who on multiple occasions helped rephrase my research in an attempt to circulate it to wider audiences. Her words of encouragement and her positive feedback have been an invaluable source of constant support. It was a pleasure collaborating with her during the last phase of my PhD.

I am particularly grateful to my husband Muhammad Atif Qureshi who never complained when I failed to cook dinner on time, and always offered to help with various chores to the best of his abilities even if that help meant burning food he tried to cook. He is a strong believer of women empowerment not just in words but in action. I am also very grateful to my daughter Fareeha Qureshi who will only be 5 months old at the

time of submission of this thesis but her lovely smile keeps me going and pushing hard.

I am highly grateful to my father who at the time of this writing is unable to walk on account of a recent stroke; his words that one day I will make him proud are what have been a motivation for me throughout this tough journey. I am also very grateful to my mother whose prayers have stood by me and helped me overcome all encountered difficulties.

Lastly, and most importantly, I am humbled by the Blessings of the Creator and Sustainer of the Universe, Allah swt. Indeed, it is He who grants us what we do not deserve and none is worthy of praise except Him.

Abstract

Personalization efforts aim to alleviate the “information overload” problem in an attempt to help users address their information needs in the best way possible. An increasing number of systems that employ personalization have cropped up in recent past with even well-known commercial giants targeting their efforts towards enhanced personalization within their services e.g. Amazon product recommendations, Netflix movie recommendations, Google Now etc. A fundamental building block of any personalization attempt is the user model that powers it. User modelling has remained a theme central within the broad research area of personalization with most traditional sources for user modelling being controversial in nature on account of the loss of privacy associated with them. With the advent of the Social Web, a paradigm shift has occurred in the way content is generated on the Web leading it to become an online gathering point for the masses. Users now leave traces of their online experiences on various Social Web platforms referred to as “social breadcrumbs” in the context of this thesis. Recent research efforts began to explore the possibility of utilizing Social Web data for creation of personalization-centric user models; most of the approaches attempted to make use of bookmarks and social tags for user modelling. These sources however are less effective on account of few users making use of bookmarking and social annotation tools rendering them infeasible for large-scale application in personalized applications.

Given the limitations of current user modelling efforts, we explored social network usage patterns and personalization-related privacy concerns in an attempt to derive aspects of Social Web data that can lead towards effective user profiles. The analyzed correlations led us towards the proposition of a Twitter-based user model which takes into account not only the language usage patterns of the user under consideration but also users in his/her network. More specifically, a framework based on

statistical language models is proposed. This model enables us to model the probability distribution of words within a user's language that he/she employs over Twitter in addition to the probability distribution of words within those user's language whom he considers trustworthy (on Twitter). The expressive nature of the user modelling efforts are depicted via the incorporation of two similarity measures into the model whereby common users within a network are utilized for the network-based similarity measure, and common topical interests are utilized within the topical similarity measures. To the best of our knowledge, this work constitutes one of the first attempts to take into account social network usage information for the generation of user profiles.

The proposed model was extensively explored in the context of two application scenarios, namely Web search personalization and scientific articles' recommendation, and both of these are fundamentally quite challenging in nature. For application to Web search personalization, we take into account various Twitter behaviors a user engages in. Adjustment of the parameters on basis of the Twitter behavior-based heuristics demonstrate an effective solution to personalized Web search which was verified via extensive offline and online experimental evaluations. Similarly, for application to scientific articles' recommendation the model was adjusted by only taking into account network of followed users, and replacing similarity measures with a topic modelling-based filtering measure that helps topics relevant to a user's research interest. The recommendation framework outperforms a standard baseline and produces rich recommendations of scientific articles for the user.

Contents

1 Introduction **1**

1.1 Motivation and Problem Statement . . . 1

1.1.1 Traditional Methods for User Profile Creation . . . 1

1.1.2 User Profiles Gathered from Social Web . . . 2

1.2 Open Issues . . . 3

1.2.1 Research Questions . . . 4

1.3 Contributions . . . 6

1.4 Structure of the Thesis . . . 7

2 Background **9**

2.1 Twitter Microblog Network . . . 9

2.2 Language Models . . . 11

2.3 Topic Models for Twitter Data . . . 13

2.4 Personalization on the Web . . . 14

2.4.1 Personalized Information Retrieval . . . 14

2.4.1.1 Information Retrieval Models . . . 15

2.4.1.2 Personalized Approaches to Information Retrieval . . 16

2.4.2 Recommendation Systems . . . 18

2.4.3 Evaluation Measures . . . 19

2.4.3.1 Precision and Precision@k . . . 19

2.4.3.2 Mean Average Precision . . . 20

2.4.3.3 Mean Reciprocal Rank . . . 20

2.4.3.4 Discounted Cumulative Gain (nDCG) . . . 21

2.4.4 Evaluations for Personalized Applications . . . 22

2.4.4.1 Custom Dataset for Evaluation of Personalization Efforts . . . 22

2.4.4.2 Online Interleaved Evaluation . . . 23

2.5 Summary . . . 24

3 Related Work **25**
3.1 Privacy-Personalization Paradox 25
3.1.1 Studying Users' Privacy Concerns 25
3.1.2 Privacy-Preserving Personalization 28
3.2 User Modelling 29
3.2.1 Acquisition of User-Related Information 29
3.2.2 Representation of User-Related Information 31
3.2.3 Application of the User Model 32
3.2.4 User Modelling Approaches in Lieu of Thesis Contributions . . 33
3.3 Utilization of Social Web Data for User Modelling 34
3.3.1 User Modelling through Social Tagging/Bookmarking Applications 35
3.3.2 Twitter-based User Modelling 36
3.4 Personalized Recommendation of Scholarly Works 37
3.4.1 Scientific Articles' Recommendation 37
3.4.2 Analysis of Researchers on Twitter 38
3.5 Summary 38

4 Predictors of Users' Willingness for Web Search Personalization **40**
4.1 Overview 41
4.2 Survey and Survey Results 43
4.2.1 Participants and Survey Content 43
4.2.2 Measures and Variables 44
4.2.2.1 Personalization Response 44
4.2.2.2 Facebook Usage 44
4.2.2.3 Twitter Usage 45
4.2.2.4 Q & A Activity on SNS Tools 45
4.3 Analyses and Findings 46
4.4 Prediction Model for Web Search Personalization Willingness 49
4.5 Discussion 51
4.5.1 Limitations 51
4.5.2 Implications 51
4.6 Conclusion 52
4.7 Summary 53

5 Modelling User Interests through Twitter 54

5.1 Model Overview 54

5.1.1 Novelty 55

5.2 Details of Twitter-based User Modelling Strategy 56

5.2.1 Twitter-specific behaviors 57

5.2.2 Individual Twitter models that constitute the overall model . 58

5.2.3 Similarity measures between users 61

5.2.3.1 Network-Based Similarity 61

5.2.3.2 Topical Similarity 62

5.3 Summary 63

6 Application of Twitter-based User Model in Web Search Personalization 64

6.1 Overall Architecture of Personalized Search System 64

6.2 Overall Model for Web Search Personalization: Re-ranking Framework 66

6.2.1 Using Twitter Behavior Heuristics for Parameter Setting . . . 67

6.2.1.1 Random Surfer Behavior on Twitter Network 67

6.2.1.2 Trust Scores based on Tweeting Activities 68

6.3 Evaluation of the Personalized Search Approach 68

6.3.1 Experimental Setup 69

6.3.1.1 Offline Evaluations 69

6.3.1.2 Online Evaluation 71

6.3.2 Experimental Results 71

6.3.2.1 Experiment Type I 72

6.3.2.2 Experiment Type II 73

6.3.2.3 Effect of Individual Twitter Networks 75

6.3.2.4 Online Evaluation 75

6.4 Summary 76

7 Application of Twitter-based User Model in Scientific Articles' Recommendation 77

7.1 Twitter Usage by Researchers 77

7.2 Overall Architecture of Scientific Articles' Recommendation 78

7.3 Filtering Topical Interests corresponding to Paper Titles 80

7.3.1 Ranking Module for Researchers Followed on Twitter 80

7.3.2 Ranking Module for Research Topics 81

7.4 Scoring Framework using Language Models of Followed Researchers . 82

7.5 Experimental Evaluations 84
7.5.1 Experimental Setup 84
7.5.1.1 Dataset 84
7.5.1.2 Parameters and Evaluation Measures 84
7.5.2 Experimental Results 85
7.6 Summary 87

8 Conclusion 88
8.1 Summary of Contributions 88
8.2 Answers to Research Questions 91
8.3 Significance of Research Outcome 93
8.4 Future Directions 93

A Survey 95

B Publications 99

Bibliography 101

List of Figures

2.1 Number of monthly active Twitter users worldwide from 1st quarter 2010 to 1st quarter 2015 (in millions) 10

2.2 Illustration of language model 12

2.3 Architecture of an information retrieval system 15

2.4 Architecture of a personalized information retrieval system 17

2.5 Personalized recommendations offered to a user on the Netflix movie recommendation website 18

2.6 Algorithm for Online Interleaved Evaluation borrowed from [105] 23

4.1 Overall classification results for Various Feature Types 50

5.1 Mention and Retweet by Target User 56

5.2 Twitter Biography of Target User, User Mentioned by Target User and User Retweeted by Target User (from left to right) 57

5.3 Users Followed by the Target User 58

6.1 Overall Architecture of Personalized Search System 65

7.1 Overall Architecture of Recommendation System for Scientific Articles . . 79

List of Tables

1.1 Top 20 Websites Globally 2

1.2 Difference between traditional methods and Social Web based methods for user profile creation 3

2.1 Information Retrieval Models 16

3.1 Statistics about users' privacy concerns with respect to personal data 26

3.2 Statistics about users' privacy concerns with respect to user tracking and cookies 27

3.3 Type of information users willing/unwilling to share 27

4.1 Demographic Variables (n=380) 43

4.2 Statistics for SNS Accounts of Survey Respondents 44

4.3 Statistics for Highly Used SNS Accounts by Survey Respondents . . . 44

4.4 Statistics on Users' Desirability for Web Search Personalization . . . 45

4.5 Statistics on Facebook Usage 45

4.6 Statistics on Social Network Q & A Activity 46

4.7 Logit regression showing the odds of users' willingness towards Web search personalization 47

4.8 OLS regression showing the level of users' agreement towards Web search personalization 48

6.1 Summary of Experimental Settings for Offline Evaluations 70

6.2 Statistics for Users in First Set of Offline Experiments 71

6.3 Statistics for Users in Second Set of Offline Experiments 71

6.4 Comparison of Retrieval Performance for Proposed Personalization Model against Non-Personalized Baseline and across Various Settings for *"bif"* 72

6.5 Comparison of Retrieval Performance for Proposed Personalization Model across Uniform and Trust Scores-Based Parameter Assignment and across Network and Topical Similarities 74
6.6 Comparison of Retrieval Performance for Proposed Personalization Model across Uniform and Trust Scores-Based Parameter Assignment and across Network and Topical Similarities against other Personalization Models 74

7.1 Example to Illustrate Ranking of Topics Related to Titles of Scientific Articles 82
7.2 Final Scores Assigned to Each Topic from within $Topics_p$ 82
7.3 Statistics about Employed Twitter Dataset 84
7.4 Comparison of Retrieval Performance for our Proposed Personalization Model 86
7.5 Top-7 Words from Top-5 Topics for a User in Experimental Evaluations 86

Chapter 1

Introduction

1.1 Motivation and Problem Statement

The deluge of information on the World Wide Web has given birth to the problem of *information overload* causing the "information available to exceed the user's ability to process it [25]". A fundamental solution to this problem has been a plethora of personalized applications that tailor their functionality to the needs of users [34]. The demand for personalization on the World Wide Web is growing at an unprecedented scale. The growing popularity of personalized applications such as Amazon and Netflix bear testimony to this. Effective personalization is largely dependent on the nature of models that capture users' interests and preferences to create what is known as a "user profile."

1.1.1 Traditional Methods for User Profile Creation

Traditional approaches for user profile construction make use of a user's search history (e.g., query logs and clickthrough data), browsing history, desktop documents, and/or emails [57, 169, 105]. The underlying intuition behind the use of such data is the fact that it incorporates feedback from the user[1] thereby gathering user intent through acquisition of additional contextual information from the user [135]. However, traditional methods suffer from the following major drawbacks:

- **Privacy:** The nature of the data utilized in traditional models may be too personal or sensitive and users are particularly cautious about its use [93].

- **Noise:** There is huge amount of noise in traditional data sources [112] due to the uncertainty of users' information need.

[1] History data in particular is an effective form of user feedback.

Rank	Site	Rank	Site
1	google.com	11	google.co.in
2	**facebook.com**	12	live.com
3	**youtube.com**	13	sina.com.cn
4	yahoo.com	14	**weibo.com**
5	baidu.com	15	**linkedin.com**
6	amazon.com	16	yahoo.co.jp
7	**wikipedia.org**	17	google.co.jp
8	taobao.com	18	ebay.com
9	**twitter.com**	19	yandex.ru
10	**qq.com**	20	tmail.com

Table 1.1: Top 20 Websites Globally

- **Lack of Networking Context:** The World Wide Web has transitioned from being merely a “Web of Documents” to incorporate a “Web of People” [22]; however, information about the network activity of users is not available in traditional sources.

1.1.2 User Profiles Gathered from Social Web

The World Wide Web has turned into an online gathering point for the masses, and this evolution is now referred to as the Social Web. This change enables users to leave traces of their social experiences which we refer to as social breadcrumbs[2]. In recent years, the Social Web has enabled a new form of user collaboration where users engage within a social network while at the same time generating their own content which is popularly known as user-generated content [183]. The Social Web affords users with the opportunity to share, communicate, connect, interact, and create user-generated data at an unprecedented rate. Significant constituents of the Social Web are popular social media platforms such as Facebook[3], LinkedIn[4], Twitter[5], MySpace[6], Wikipedia[7], and YouTube[8]. Table 1.1 shows seven social media sites (marked in bold) to be among the top 20 websites ranked globally according to usage (Internet traffic by Alexa on June 5, 2015).

[2] Simply, footprints on the Social Web.

[3] http://www.facebook.com

[4] http://www.linkedin.com

[5] http://www.twitter.com

[6] http://www.myspace.com

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

[8] http://www.youtube.com

The growing use of social media applications allows users to leave behind social breadcrumbs from which their interests and preferences can be automatically extracted and this has become an essential part of the user profiling process for personalized applications. Given the fact that social media contains significant pointers related to users' interests and preferences in the form of clear and explicit topics, an increasing number of research is making use of it for the extraction, analysis and representation of users' interests [29, 39, 74, 87]. In summary, Table 1.2 lists the significant differences between use of traditional methods and methods that create profiles from the Social Web.

Table 1.2: Difference between traditional methods and Social Web based methods for user profile creation

Traditional Methods	**Methods using Social Web Data**
Data is of a private nature and, hence not easily available.	Data is publicly available and, users do not have privacy concerns while sharing it.
Data is implicit in nature.	Data contains explicit indication of user interests.
Data contains a lot of noisy user interactions.	Data contains clear mentions of topics that are of interest to the user.
Data does not contain information about network connections.	Data being derived from social networks contains rich information about those whom user trusts.

1.2 Open Issues

Existing approaches that make use of social media data for personalization are fundamentally built upon the following ideas:

- Making use of Web page social annotations (folksonomies)[9] for improved personalization [21]

[9] These are mostly available on sites like Delicious, CiteULike, Flickr

- Making use of social media data to identify trending topics (e.g. news items) for improved ranking/recommendation of Web items centered on the trending topics [5, 55]

- Making use of friendship links within the social network of a user to filter the information presented to the user [39]

There remain a number of unaddressed issues with respect to the above-mentioned ideas:

- Social annotations of Web pages (folksonomies) are useful for only a minor portion of the Web and mostly cover news articles. Moreover, due to the ease (i.e., openness) of social tagging services, social annotations are highly prone to social spammers [149].

- Trending topics on social media are mostly composed of news articles and leveraging this data has a greater chance of improving personalization for news articles thereby ignoring general-purpose personalization.

- Despite the effectiveness and novelty that comes through utilization of a user's social network friendship links this approach remains limited to an enterprise setting due to the ease of availability of such data being limited to enterprises.

Lastly and most significantly, understanding the target audience of personalized applications is an important aspect for the development of meaningful and well-accepted applications. In particular users' privacy concerns have proven to be a significant challenge with respect to personalization, and the issue of when to personalize and when not personalize represents an important challenge, which has not been deeply investigated yet with the exception of some basic analysis [173].

1.2.1 Research Questions

In the following chapters the thesis will address and answer several research questions which are summarized in this section. With the above-stated open issues in mind, we can ask the following question, which is the core research question of this thesis:

How can we effectively utilize social breadcrumbs left behind by users to reflect their interests and preferences for improved personalization?

Stemming from the above core research question are the following specific research questions:

- **Can a user's social network usage patterns serve as a window into his/her privacy concerns with respect to personalization?**
  As a motivating example, let us consider a scenario involving two users: user A is highly active on various social networking services as he/she communicates his/her thoughts over a range of topics (politics, religion, economics etc.), while user B is much less active when compared to user A in sharing his/her thoughts on social networking services. In line with this scenario it would be interesting to investigate the correlations of the behaviors of both users A and B and their openness to personalization. Some commercial systems have attempted a similar integration (as is evident in recent social search approaches by Google and Bing: e.g., Bing's Facebook integration[10] and Google's "Search Plus Your World" Google+ integration[11]). However, to the best of our knowledge, there is no literature that describes the correlations between social network usage patterns and privacy concerns.

- **Can sources other than tags and Web page annotations be utilized as a source of evidence for user profile creation?**
  As mentioned in Section 1.2, earlier research efforts that aim to exploit information from the Social Web for personalized applications rely mostly on social bookmarking and tagging systems [125, 179]. However, Heymann et al. [79] questioned the usability of bookmarking meta-data for Web search engines by collecting a very large dataset (in fact, the largest known to the academic community) from a social bookmarking site. Heymann et al.'s findings revealed that social bookmarking lacks the size and distribution of tags necessary to make a significant impact for information retrieval at large. Hence, we explore microblogs and in particular Twitter as an alternate source of user profile creation.

- **Is it possible to utilize the social network information (friendship links) of a user in order to create a richer and more enhanced user profile?**
  As mentioned in Table 1.2, traditional data sources lack detailed information

[10] http://www.bing.com/community/site_blogs/b/search/archive/2011/05/16/news-announcement-may-17.aspx

[11] http://www.google.com/insidesearch/plus.html

about the users that a particular user trusts and hence, current models are unable to take into account the interests and preferences of these trustworthy users. We attempt to alleviate this shortcoming within current research efforts by exploiting aspects of a user's Twitter network (users followed, mentioned, retweeted) in a language modelling framework to generate a user model while also incorporating similarity measures between a user and his/her network and trust scores based upon user interactions on Twitter.

1.3 Contributions

As a case-study for our thesis on user profiling for personalized applications, we take up personalized Web search as the example application in two contributions while a recommendation system for scientific articles in the third contribution. With the help of these case-studies, following are the main contributions stemming from this thesis:

- Using data gathered in a user survey, we present an analysis of the correlation between the users' willingness to personalize Web search and their social network usage patterns [199]. The participants' responses to the survey questions enabled us to use a regression model for identifying the relationship between SNS variables and willingness to personalize Web search. We also performed a follow-up user survey for use in a support vector machine (SVM) based prediction framework. The prediction results lead to the observation that SNS features such as a user's demographic factors (such as age, gender, location), a user's presence or absence on Twitter and Google+, amount of activity on Twitter and Google+ along with the user's tendency to ask questions on social networks are significant predictors in characterising users who would be willing to opt for personalized Web search results.

- This thesis explores the use of the Twitter microblog network as a source of user profile construction for personalized Web search [200] and scientific articles' recommendation [201]. To the best of our knowledge, the use of microblogging platforms and, in particular, Twitter has not, with the exception of a few works [7, 87], been explored as a source of user profile construction for personalized applications and we undertake such a direction in this work.

- This thesis proposes a novel approach utilizing statistical language models that helps in capturing users' interests and preferences; the approach models a user's language and related interests through utilization of their microblog (i.e., Twitter) generation and usage patterns [200, 201].

- This thesis proposes a model for user profiling that utilizes a user's microblog (i.e., Twitter) network for tapping into interests and preferences of the user under consideration [200, 201]. In particular, aspects of a user's Twitter network (users followed, mentioned, retweeted) are used in a language modelling framework to generate a user model while also incorporating similarity measures between a user and his/her network and trust scores based upon user interactions on Twitter.

1.4 Structure of the Thesis

This dissertation is structured as follows. In this Chapter, we introduced the research motivations and problem statement along with a presentation of the research questions and contributions. Chapter 2 presents some background material to provide a description of information retrieval, recommendation systems along with a description of some privacy concerns with respect to personalization on the Web. Chapter 3 presents a description of the state-of-the-art related to the core research areas of the thesis. It introduces the most relevant definitions and the related work for the research fields of user modelling, Web search personalization, social recommendation algorithms, and scientific articles' recommendation. Chapters from 4 to 7 are the core Chapters of the thesis and include our main contributions.

Chapter 4 presents our study on the correlation between users' social network usage patterns and their privacy concerns with respect to Web search personalization. We conducted a large-scale user survey in two parts where the first part gathered responses from 380 people from various countries, and the second part gathered responses from 113 people from various countries. This data was then used in a regression model to analyse the correlation between SNS variables and willingness to personalize Web search. The data was also used in a support vector machine (SVM) model to explore the potential to make predictions about users who would be willing to opt for personalized Web search results. We wish to explore if the prediction accuracy would be sufficient for a real personalised Web search system.

Chapter 5 presents our proposed user profiling model that we build through utilization of a statistical language modelling approach over a user's Twitter data. We explain the individual constituents of the model followed by an explanation of its strengths and limitations.

Chapter 6 presents details of application of our model to personalized Web search. Through experimental evaluations, we demonstrate the effectiveness of the model and how it beats other approaches in the literature.

Chapter 7 presents details of application of our model to scientific articles' recommendation. Through experimental evaluations, we demonstrate the usefulness of the proposed recommendations.

Chapter 8 concludes this thesis with a discussion on findings and an outline of future work.

Chapter 2

Background

In order to cover the related background issues, we start by giving a review of the microblogging platform Twitter which is our main source for extraction of social breadcrumbs. This is followed by an overview of language modelling which is the essential approach for our proposed user profiling model, and a brief overview of topic modelling which we use as a similarity measure within our proposed model. Finally, we conclude this chapter with an overview of some personalized Web applications particularly stressing upon personalized Web search and recommendation systems along with a brief overview of evaluation measures and techniques used to assess the quality of personalization algorithms.

2.1 Twitter Microblog Network

The past few years have seen a huge growth in the use of microblogging services and as of today these services constitute an essential part of the Social Web. Microblogs are what we can term as "real-time blog services that allow users to post short messages" and the most popular microblogging platform to date is Twitter. Figure 2.1 shows the quarterly growth of active Twitter users since 2010[1]. The unique aspect of Twitter is that it allows users to indulge in "social networking" along with "microblogging" [97]. The Twitter microblog network enables a user to follow any other user and unlike most online social networking sites the relationships of *following* and *being followed* require no reciprocation; and hence, it is primarily an "interest" based social network [152]. The various features of Twitter provide an opportunity for the application of algorithms for using Twitter data. Below we present an overview of the features of Twitter to aid the understanding of the reader in subsequent chapters:

[1] statista.com is the primary source of these statistics.

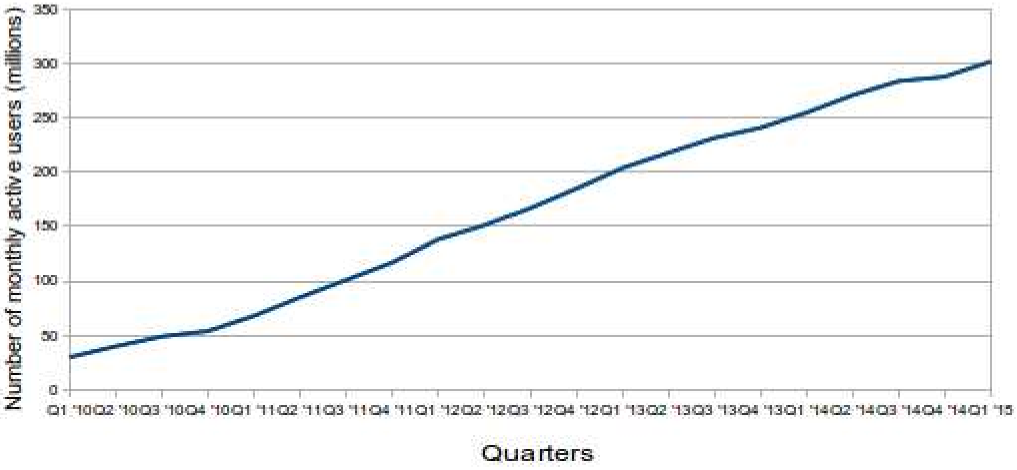


Figure 2.1: Number of monthly active Twitter users worldwide from 1st quarter 2010 to 1st quarter 2015 (in millions)

- ***Tweet:*** The 140-character message that users can post on the web site is called a "tweet". When a user posts a tweet, it is displayed on the user's timeline while also viewable by all of his/her followers. Moreover, the tweet may also be viewed by any other Twitter user searching for keywords matching some content in the tweet.

- ***Mention:*** The feature of Twitter that enables Twitter users to engage in conversations with each other is called "Mention". The process requires use of '@' followed by the user identifier (name) in the tweets e.g., *`Guys what's the plan for today? @user1 @user2`*.

- ***Retweet:*** The feature of Twitter that allows a Twitter user to share a tweet he/she comes across to his/her entire network of followers is called "Retweet." This feature of Twitter is one of most innovative features ever invented by a social media platform and lead Facebook to introduce a similar "Share" button.

- ***Lists:*** A feature of Twitter that allows grouping of users with similar interests is called "Lists."

- ***Hashtags:*** Any keyword(s) preceded by the '#' symbol is used for annotation of tweets making it easy to search, track and follow. Such keywords are called hashtags.

Our main motivations behind the use of Twitter as a source of user profile construction are as follows:

- Tweets are publicly visible to everyone on the web by default; however, Twitter users can apply privacy to their Twitter accounts by making their accounts *protected*. This alleviates privacy concerns to a large extent by giving users explicit control over their privacy settings.
- Twitter provides an important platform for people to express opinions, share ideas, receive updates relating to various topics of interest (e.g., science, sports, politics), and discover latest news. Hence, it has become an avenue where people explicitly express their interests and preferences which can be mined for creation of user profiles [53].
- Data acquired from users' search history and browser history is sparse [138] particularly for new users. Twitter users however follow other users who reflect their interests. This implies that the profile of a given user can be created through his/her followed users (i.e., the network of users to whom he/she is connected) even if the given user himself/herself does not post much content.
- Twitter provides an easy-to-use API[2] for getting its data whereas history data (both search history data and browser history data) is mostly available in commercial settings. The API enables researchers to gather not only tweet content but also network information for Twitter users.

2.2 Language Models

A language model is a statistical model that assigns probabilities to sequences of words which appears in a sentence thus forming a style of specific vocabulary usage for a language [148]. Language models have been used in different natural language processing applications such as speech recognition, information retrieval, handwriting recognition, spell correction, etc.

Consider a language where V is the set of all words or entire vocabulary of a language i.e., {apple, the, happy, cat, ...}. A sentence can be formed by using these words in a sequence, and by observing different sentences we can generate a probability distribution of these words i.e., which word is likely to appear more frequently or which word is likely to appear more frequently when surrounded by some sequence

[2] https://dev.twitter.com/

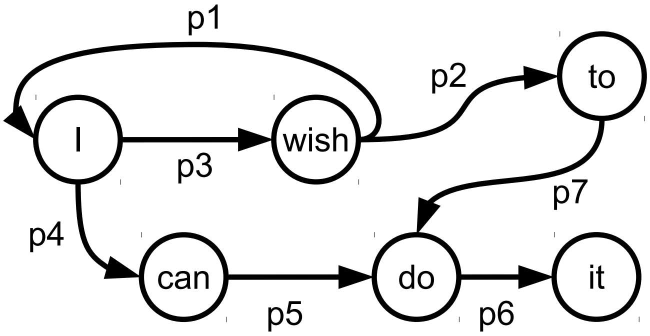


Figure 2.2: Illustration of language model

of words. Therefore, once we learned these probabilities from a training set we can conclude the probability of unseen sentences such as P_{LM}*(this is a book)* > P_{LM}*(this a book is).* Figure 2.2 shows a simple example of how a language model works, and it can be seen that P_{LM} *(I can)* > P_{LM}*(I it)* mainly because "I" is commonly followed by "can" within natural language[3]. It should also be noted that in what follows we will not use models in which the probabilities depend upon pairs of words, or from the position of the word within a sentence.

The following are few examples of the application of language models:

- In speech recognition, words with similar sounds can be corrected by utilizing a trained language model obtained through observation of the surrounding context of words.

- In spelling correction, misspelled words can be corrected by exploiting the surrounding words in context by means of a language model over characters/alphabets.

- The authorship problem (i.e., identifying authors of a passage of text) can be solved if we train a language model each for different authors and then by utilization of the trained language model we can identify authorship of a disputed literary piece among the known authors. For example, if we train a language model using the literature produced by William Shakespeare and we also train another language model using the literature produced by Jane Austen, then using these models we can predict the authorship of a literary piece produced by either of the authors when the authorship is disputed between these two.

[3] Note that corpora of natural language are used for generation of probabilities for language models.

- In information retrieval, a language model is trained for each document in the collection and the ranking function of the retrieval system ranks documents based on the probability of the query in the document's language model [136].

2.3 Topic Models for Twitter Data

Topic models are a family of statistical models that were designed to aid in text summarization through extraction of latent topics within the texts [162]. The most well-known of these is the one proposed by Blei et al. i.e. Latent Dirichlet Allocation (LDA) [27].

The fundamental ideas behind topic models are quite simple and are based on a probabilistic model for each document in a given text collection where topic is represented as a multinomial probability distribution over V unique words in the corpus vocabulary. More specifically a topic is defined in terms of a probability vector $p(w \mid t) = [p(w_1 \mid t), . . . , p(w_V \mid t)]$, where $\sum_v p(w_v \mid t) = 1$ and there are T topics in total. A document is represented as a finite mixture of the T topics, and there is a total of N documents. Each document d is assumed to have its own set of mixture coefficients $[p(t = 1 \mid d), . . . , p(t = T \mid d)]$ i.e., a multinomial probability vector such that $\sum_t p(t \mid d) = 1$. Hence, a randomly selected word from document d has a conditional distribution $p(w \mid d)$ that is a mixture over topics, where each topic is a multinomial over words:

$$p(w \mid d) = \sum_{t=1}^{T} p(w \mid t)p(t \mid d)$$

The recent past has seen research efforts being devoted to development of a variety of extensions to topic models [106, 142, 147] with few topic models devoted specifically to Twitter [109, 139, 141, 202]. We use the model proposed by Zhou et al. namely Twitter-LDA [202]. Twitter-LDA is fundamentally an author-topic model implying that it is a generative, probabilistic model for authors and documents wherein each author is associated with a mixture over topics. Note that this is appropriate to the modelling criteria within this thesis in that each Twitter user is interested in a specific set of topics (we revisit this notion in Chapter 5 during explanation of similarity measures based on Twitter-LDA). More specifically, Twitter-LDA is built upon following assumptions:

- There is a collection of 'K' topics in Twitter with each topic represented by a word distribution.
- Each user's interests are modeled through a distribution over topics.
- When writing a tweet, a user may choose to write a background word or he/she may choose a topic based on his topic distribution which may then lead to the choice of a word based on the word distribution of the chosen topic.

Twitter-LDA differs from the original LDA framework by Blei et al. [27] in that a single tweet is assigned a single topic instead of a distribution over topics. This is more suited to the tasks considered in the context of this thesis that utilize Twitter data for various aspects within user profile creation.

2.4 Personalization on the Web

Montgomery and Smith [116] provide a definition of personalization from a marketing perspective, "the adaptation of products and services by the producer for the consumer using information that has been inferred from the consumer's behavior or transactions." With regards to personalization on the Web, Blom [120] refers to it as, "A process that changes the functionality, interface, information content, or distinctiveness of a system to increase its personal relevance to an individual." Moreover, Web-based personalization provides an effective solution to the "information overload" [25, 120] problem that exists due to the huge amount of information available on the World Wide Web.

Currently, the popular constituents of current personalization systems on the Web are personalized information retrieval systems [69] and recommendation systems [88]. The next two subsections present an overview of each.

2.4.1 Personalized Information Retrieval

The process through which users' information needs are satisfied is termed as information retrieval. As a research field, it is primarily concerned with representation, storage, organization of, and access to a collection of text documents (containing information) in order to help a user find "relevant information" [20, 150]. The notion of "relevance" is heavily tied to an information need which is usually motivated by a real-world task and expressed in the form of a "textual query". As an example, a

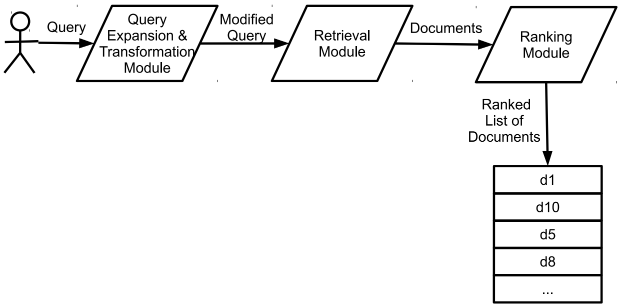


Figure 2.3: Architecture of an information retrieval system

Computer Science student in making a decision between using the programming language *Java* or the programming language *Python* for a particular programming task, may think of *"Java vs. Python speed"* as the query to express his/her information need. The challenge for information retrieval systems lies in correct interpretation of the user's information need through few query terms.

Figure 2.3 shows the architecture of a general-purpose information retrieval system. A user issues the query in the form of keywords, and this query is fed to the query processing module where the query is transformed into a set of index terms. The modified query is processed by the retrieval module which returns a list of retrieved documents that contain at least one term of the modified query. This list includes candidate relevant documents with respect to the query. Finally, the ranking module processes the list of candidate relevant documents in order to assign a relevance score to each document; this relevance score reflects the similarity between the query and document. The final output i.e., the ranked list of documents contains documents in decreasing order of their relevance scores and this list is presented to the user.

2.4.1.1 Information Retrieval Models

An information retrieval model is a conceptual framework for representation of documents and queries as well as the definition of a ranking framework for retrieved documents. The simplest model for query and document representation within documents and queries is the bag-of-words model which identifies each term as a single word thereby ignoring term order. Based on the bag-of-words representation, three different types of information retrieval models for document ranking have been proposed and the difference between these model types lies in the underlying mathematical framework. Table 2.4.1.1 provides information about these models; for the purpose of this thesis we have utilized probabilistic models and we briefly present one model here (i.e., BM25) which we use as our non-personalized baseline in Chapter 6.

Table 2.1: Information Retrieval Models

Model	Representation Details	Operations Performed
Boolean	Documents and queries represented as set of terms	Set-based operations
Vector Space	Documents and queries represented as vectors in multi-dimensional space	Algebraic operations
Probabilistic	Documents and queries represented as probability distributions	Probabilistic operations

The BM25 model, proposed by Robertson and Walker, was first introduced at TREC in 1995 [145, 146]. The ranking function is given by

$$P(R=1 \mid d) \approx \sum_{w \in Q \cap d} tf_{w,Q} \frac{(k_1+1)tf_{w,d}}{k_1((1-b)+b\frac{|d|}{|d|_{avg}})+tf_{w,d}} log \frac{N-df_w+0.5}{df_w+0.5}$$

where $tf_{w,Q}$ is the number of times term w appears in the query, $tf_{w,d}$ is the number of times term w appears in the document, $|d|$ is the number of terms in the document, $|d|_{avg}$ is the average document length, N is the number of documents in the collection, df_w is the number of documents in which term w occurs in the collection and k_1 and b are model parameters.

2.4.1.2 Personalized Approaches to Information Retrieval

One characteristic of general-purpose information retrieval (IR) systems is that the systems return the same results to different users submitting the same query. However, there is a huge amount of diversity in the information needs of users. To understand the concept of diversity in information needs, recall the previous example from this section of a Computer Science student requiring information about programming languages; a query "python" in this context would yield search results containing documents on the snake and on the programming language. To correctly satisfy the information need in this setting, the information retrieval system would have to take the user (i.e., the Computer Science student) into account [35] since the

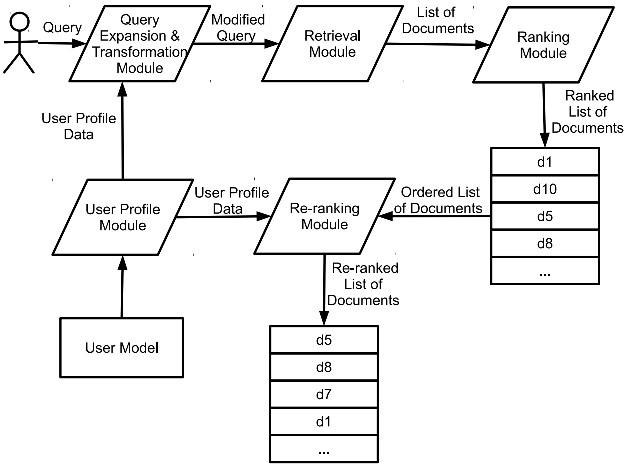


Figure 2.4: Architecture of a personalized information retrieval system

relevant documents for him/her would be documents on the programming language "python". Recent years have seen the emergence of personalized approaches information retrieval as an effective approach to deal with such diversity in users' information needs [172, 159, 191].

Figure 2.4 shows the architecture of a personalized information retrieval system. The personalized information retrieval system differs from the general-purpose information retrieval system in that the personalized system contains a user profile module that utilizes a user model containing user's personal information, interests and preferences. As can be seen from the Figure, the user profile module can play a role in two ways:

- ***Query Adaptation:*** This technique comprises enrichment of the user's submitted query terms in order to arrive at a better representation of the user's information needs. The enrichment is performed with the help of the user model [45, 204] as shown in Figure 2.4 where the user profile module feeds user profile data into the query expansion and transformation module.

- ***Result Adaptation:*** This technique comprises re-ranking of the original ranked list of documents. As shown in Figure 2.4, this re-ranking is performed through an additional ranking step for re-ordering of documents based on the user model [125, 160, 172, 179].

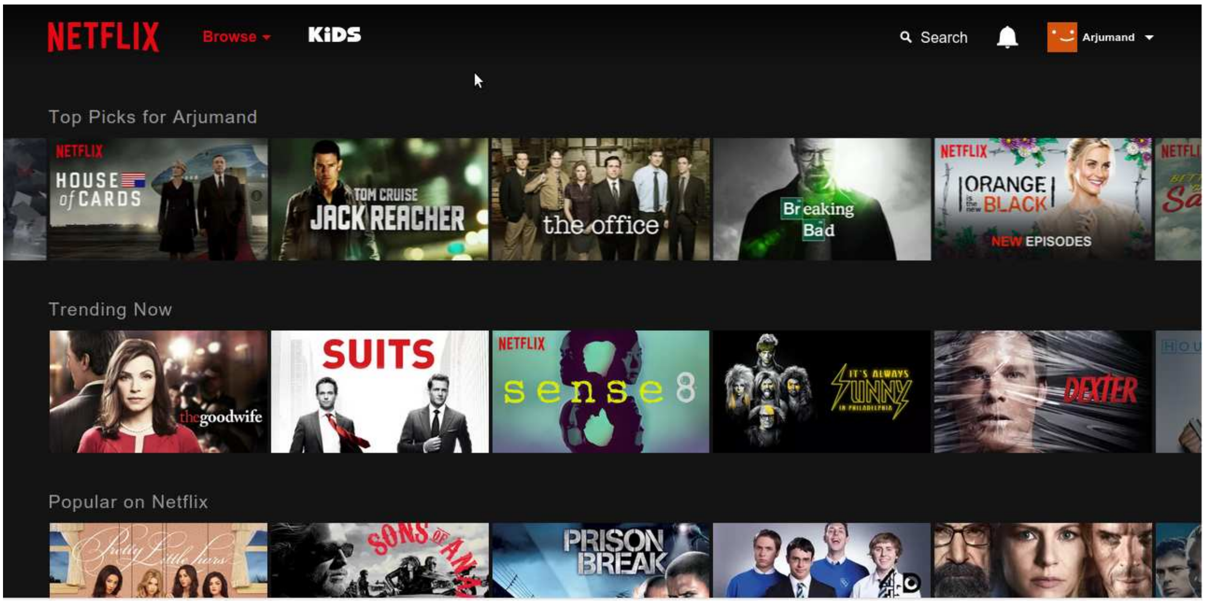


Figure 2.5: Personalized recommendations offered to a user on the Netflix movie recommendation website

2.4.2 Recommendation Systems

Recommendation systems fall under the category of information filtering systems and are primarily software systems that provide suggestions for items of interest to a user [88] in a personalized manner. Some common application scenarios for recommendation systems are movie recommendations (see Figure 2.5 for an example of a popular movie recommendation system called *Netflix*), book recommendations, news recommendations, product recommendations etc. Recommendation systems have been able to enhance the users' experience of exploring and finding new and interesting content, and this is particularly applicable for the e-commerce domain. According to the classification of recommendation algorithms by Burke [36], following are the main groups of recommendation algorithms:

- ***Collaborative Filtering:*** This class of algorithms aim towards prediction of a user's interests through exploitation of consumption patterns of users and their preferences in order to uncover similarities between users [78]. Similarity in the collaborative filtering context implies having a preference for similar items and thereby providing similar ratings for various items.

- ***Content-Based Recommendation:*** This class of algorithms utilize item descriptions (mostly textual such as item metadata, tags, genre etc.) and a user model that assigns importance to various characteristics within the item descriptions. Generally, keywords are used for item descriptions and based on user profile information of what kind of items user has liked in the past, content-based similarities are computed and through these similarities new items are recommended [130].

- ***Knowledge-based Recommendation:*** This class of algorithms exploit additional, often manually provided knowledge about items to be recommended. Such knowledge includes domain-specific characteristics of the items that constitutes information about how the item meets certain user needs and preferences [144].
- ***Hybrid Algorithms:*** This class of algorithms combines multiple techniques together to achieve some synergy between them [37]. Netflix shown in Figure 2.5 utilizes hybrid algorithms to make movie recommendations.

2.4.3 Evaluation Measures

Evaluation is the process of systematically quantifying the results produced by a personalized system in order to measure the degree to which it achieves user satisfaction. This measurement is performed via calculation of a quantitative metric which is directly associated with the relevance of the results to the user. A common approach to compute such a metric is to compare the results produced by the system with results suggested by humans corresponding to a certain information need. In the following, we define five evaluation metrics used throughout this thesis; these evaluation metrics are extensively explained in the book "Modern Information Retrieval" [20].

2.4.3.1 Precision and Precision@k

Consider a set of documents relevant to the user D_r[4] and a set of documents generated by a given personalization system D_a. Precision is then defined as follows:

$$Precision = p = \frac{|D_r \cap D_a|}{|D_a|}$$

[4] Note that the set of relevant documents contains explicit relevance judgements from the users themselves.

where $|D_r \cap D_a|$ is the number of documents common between sets D_r i.e., intersection of two sets.

Precision@k is defined as the ratio of relevant documents over the first top-k results. This measure is generally used when the list of returned documents is huge, i.e. when relevance judgements for all the documents retrieved cannot be assessed by manual annotators (such as returned result set in Web search engines [12]). It is defined as follows.

$$P@k = \frac{|D_r|}{|Results\ \ at\ \ top\text{-}k|}$$

2.4.3.2 Mean Average Precision

The underlying principle behind mean average precision measure is to produce a summary value of the ranking by averaging precision figures after observation of each new relevant document. Let D_{r_i} refer to the set of relevant documents for query q_i, and let $D_{r_i}[k]$ be a reference to the kth document in D_{r_i}. Then, $P(D_{r_i}[k])$ is the precision when the $D_{r_i}[k]$ document is observed in the ranking of q_i. The average precision AP_i for query q_i is defined as follows:

$$AP_i = \frac{1}{|D_{r_i}|}\sum_{k=1}^{D_{r_i}} P(D_{r_i}[k])$$

MAP, the mean average precision over a set of queries is then defined as follows:

$$MAP = \frac{1}{|Q|}\sum_{i=1}^{|Q|} AP_i$$

where Q is the set of queries.

2.4.3.3 Mean Reciprocal Rank

Mean reciprocal rank is particularly important in cases where the first relevant document is of interest to a user such as in question answering applications. This evaluation metric favors results whose first correct is higher in the ranking.

Let D_{a_i} be the ranking relative to a query q_i and let $S_{correct}(D_{a_i})$ be a function that returns the position of the first correct answer D_{a_i}. Given a threshold ranking position S_h, the reciprocal rank of D_{a_i} is defined as

$$D_{a_i} = \begin{cases} \frac{1}{S_{correct}(D_{a_i})} & if S_{correct}(D_{a_i}) \leq S_h \\ 0 & otherwise \end{cases}$$

That is, the reciprocal rank is zero if the first correct result occurs at a position in the ranking beyond S_h. For a set Q of queries, the Mean Reciprocal Rank (MRR) is the average of all reciprocal ranks, which is computed as follows:

$$MRR = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{S_{correct}(D_{a_i})}$$

2.4.3.4 Discounted Cumulative Gain (nDCG)

Discounted Cumulated Gain (DCG) is an evaluation metric that takes into account graded relevance judgements for documents (i.e. rel_i), and penalizes the evaluation of a search algorithm if it retrieves relevant documents late in the ranking. The DCG at a particular rank position p is defined as:

$$DCG_p = \sum_{i=1}^{p} \frac{2^{rel_i} - 1}{log_2(i+1)}$$

The variation in result set size for different queries and different systems implies DCG has to be normalized and to this end an ideal DCG is calculated after sorting documents of a result list by relevance in order to produce an ideal DCG at position p (i.e., $IDCG_p$). Normalized DCG is then as follows:

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

2.4.4 Evaluations for Personalized Applications

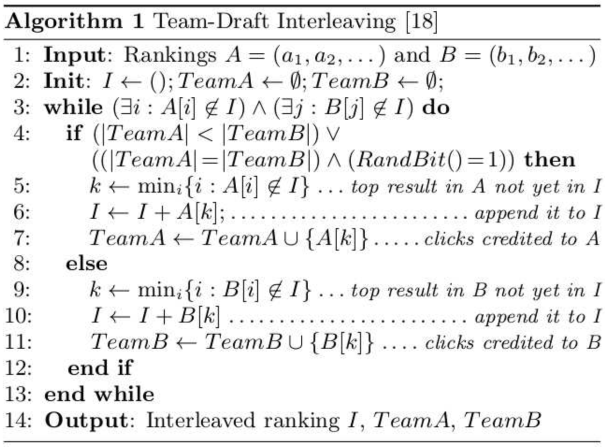
The evaluation of any personalized search algorithms is a very challenging research issue. The challenge arises due to the nature of the tasks involved in suggestion of personalized content to users [182]. The user is the central component of any personalization process and hence, there is a strong need for user-centric evaluation efforts. Traditional information retrieval systems are evaluated by means of the Cranfield paradigm [115] which ensures repeatable and controlled experiments. Such a system-centered evaluation however is not suited to personalized search systems and hence, various research efforts have been devoted to evaluation of these systems. In the following subsections we explain two such efforts that aim to alleviate issues with respect to evaluation of personalized information retrieval systems.

2.4.4.1 Custom Dataset for Evaluation of Personalization Efforts

Harpale et al. [77] were one of the first to create a custom dataset for benchmark evaluations of personalized search performance. *CiteData* is a collection of academic papers selected from CiteULike[5] social tagging web site's database and filtered through CiteSeer's[6] database for cleaning meta-data regarding each paper. It comprises 81,432 academic articles along with a rich link structure constituting references between papers. The dataset contains personalized queries and relevance feedback scores on the results of those queries. Specifically, it includes 41 queries obtained through efforts at mimicking personalized relevance feedback by gathering a group of volunteer experts (graduate and PhD students). The volunteer experts were selected with the aid of CiteULike groups by selecting those topics within the groups that fit the research areas of the volunteer annotators. Finally, the volunteer annotators were asked to design their own custom search tasks and come up with queries corresponding to those tasks thereby imitating a real-life situation wherein a Computer Science researcher is interested in finding papers according to his/her information need. The volunteers assigned relevance judgements to a set of retrieved documents obtained by pooling from seven different retrieval algorithms. The usefulness of CiteData has been established by several works [22, 168, 200] thereby proving the validity of the techniques undertaken by Harpale et al. to ensure its usefulness.

[5] `http://www.citeulike.org/`

[6] `http://www.citeulike.org/`



Algorithm 1 Team-Draft Interleaving [18]

```
Input: Rankings $A = (a_1, a_2, \dots)$ and $B = (b_1, b_2, \dots)$
Init: $I \leftarrow (); TeamA \leftarrow \emptyset; TeamB \leftarrow \emptyset;$
while $(\exists i : A[i] \notin I) \wedge (\exists j : B[j] \notin I)$ do
  if $(|TeamA| < |TeamB|) \vee$
     $((|TeamA| = |TeamB|) \wedge (RandBit() = 1))$ then
    $k \leftarrow \min_i\{i : A[i] \notin I\}$ ... top result in A not yet in I
    $I \leftarrow I + A[k];$ ........................ append it to I
    $TeamA \leftarrow TeamA \cup \{A[k]\}$ ..... clicks credited to A
  else
    $k \leftarrow \min_i\{i : B[i] \notin I\}$ ... top result in B not yet in I
    $I \leftarrow I + B[k]$ ........................ append it to I
    $TeamB \leftarrow TeamB \cup \{B[k]\}$ .... clicks credited to B
  end if
end while
Output: Interleaved ranking $I$, $TeamA$, $TeamB$
```

Figure 2.6: Algorithm for Online Interleaved Evaluation borrowed from [105]

2.4.4.2 Online Interleaved Evaluation

Matthijs and Radlinski considered online retrieval experiments for evaluation of personalized search performance with the following underlying considerations [105]:

- Offline evaluations (with relevance judgement on collection of documents) are not true reflections of user behavior in that they do not represent a real query workload.
- A slow and tedious process is involved in making relevance judgements for personalized search evaluations in that each query and user would have a separate notion of relevance.

On account of the above considerations, they proposed utilization of an interleaved evaluation technique [140] for evaluation of personalized search performance. Interleaved evaluation is a technique that is able to combine the search results of produced from two different algorithms through alternating between results from the two search rankings while omitting duplicates. The user is presented with the interleaved ranking and clicks on documents from this ranking as if it were produced from a single ranking. The ranking that contributed the most clicks over many queries and users is considered better.

Similar to Matthijs and Radlinski, we utilize the Team-Draft interleaving algorithm for the purpose of our experiments. Figure 2.6 shows the detailed steps of the algorithm; intuitively it follows the manner of selection of sports teams in friendly games: from a pool of available players (analogous to search results in rankings A and B), two captains (one for TeamA and one for TeamB) take turns picking their next preferred player from the set of remaining players, with each turn preceded by a coin toss for determining the captain who gets to pick first. For the purpose of imitating search evaluations, the selection order for players is equivalent to ranking order for search results. A click from the user mimics a voting pattern with the vote assigned to one of the two rankings (i.e., the ranking that selected a certain clicked document first). The user gives one vote per query impression to one of the two rankings; supposing user clicks a results from ranking A and b results from ranking B. If a >b, we can say that ranking A outperforms ranking B and vice versa.

2.5 Summary

This chapter covered the essential background material that can aid the reader in understanding the contributions of this thesis. We first gave an overview of the various features of the Twitter microblog network, particularly stressing upon explaining the constituents of the follower/followee phenomenon on the social networking platform. We also explained the various functionalities offered by Twitter such as the ability to post a 140-character message on the site (known as tweet), the ability to engage in conversations (known as mentions), the ability to re-post content by other Twitter users (known as retweet), the ability to form groups of users (known as lists), and the ability to annotate tweets by beginning a word with '#' symbol (known as hashtags). We then proceeded towards a brief explanation of language models and topic models. We explored how language models are used for the language generation process whereby probabilistic distributions model words in a given sequence of text; we also covered some essential applications where usage of language models has advanced state-of-the-art. Topic models were explored as generative, probabilistic models that assign a distribution of topics to documents where each topic is a distribution over words. Finally, the Chapter covered some basics of personalization on the Web wherein we gave an overview of two of the most used personalized applications i.e. personalized Web search and recommendation systems. In this context we also defined various evaluation measures and specific evaluation techniques that we will use for the purpose of evaluating our proposed user model.

Chapter 3

Related Work

This chapter presents an overview of the state-of-the-art in user profiling applications that are related to the contributions of this thesis. In particular, we consider the privacy aspects associated with personalization in addition to surveying works on personalized Web applications that aim to alleviate users' privacy concerns. This is followed by an overview of research on user modelling where we present in detail a review of the three main stages of any user modelling framework. We then present an overview of works that utilize social Web data for user modelling. Finally, the chapter concludes with an overview of systems for scientific articles' recommendation and within this context we also discuss the role of microblogs in academia.

3.1 Privacy-Personalization Paradox

Personalized applications offer an enhanced user experience due to their capability to serve users according to their tailored needs. However, this enhancement comes at the cost of user's privacy [175], and traditionally users have been known to show reluctance in disclosing personal data or allowing their system usage to be tracked [13, 41, 174, 181, 195]. This *"privacy-personalization"* paradox [19, 195] has given birth to a number of studies that aim to characterize users' privacy concerns with respect to personalization along with introduction of systems that aim to alleviate users' privacy concerns associated with personalization. We review some of these works in the following subsections.

3.1.1 Studying Users' Privacy Concerns

Numerous surveys and user studies have demonstrated a consistent phenomenon of growing privacy concerns of computer users. The study by Kobsa was the earliest that pointed out the tension between personalization and privacy [92] albeit without much impact due to lack of personalization on the Web. With the advent of the Web and thereby with Web personalization growing at an unprecedented scale, many studies were conducted between 1998 and 2003 (mostly in the United States) [174]. We summarize these studies along the following three dimensions [93]:

- Personal data
- User tracking and cookies
- Information type

Privacy Concern	**Statistics**
Internet users concerned about privacy of personal information online	89.5% [2], 83% [3]
People who have refused to provide personal information to website at least once	95% [81]
Internet users supplying false or fictitious information to a website when asked to register	40% [81], 15% more than half of the time [153]
People who are concerned if a business shares their data for a purpose different from the one for which they were originally collected	90% [1]

Table 3.1: Statistics about users' privacy concerns with respect to personal data

Tables 3.1 and 3.2 provide summary statistics with respect to personal data and use tracking and cookies respectively, while Table 3.3 provide details of information types and the level of willingness with which users share the information. Furthermore, a 2010 survey by Anton et al. [16] demonstrated increasing privacy concerns regarding website personalization between 2002 and 2008. Another survey by Turow et al. [177] discovered that 66% of Americans across all genders and age groups are opposed to marketing of personalized advertisements. Behavioral log analysis makes most users uncomfortable as this gives them the feeling of being watched and tracked

Privacy Concern	Statistics
People concerned about being tracked on the Internet	54% [64],63% [86]
People concerned that someone might know what websites they visited	31% [64]
Users who feel uncomfortable being tracked across multiple websites	91% [86].
Internet users who set their computers to reject cookies	25% [49], 10% [64]

Table 3.2: Statistics about users' privacy concerns with respect to user tracking and cookies

Type of Information	Willingness to Share
Personal preferences such as favorite televison show and favorite snack	Users share this information willingly [9, 111, 133]
Financial information such as credit card number	Users extremely protective of this information [111, 133]
Demographic and life-style information	Users share this information somewhat willingly [111, 133]
Identification information and online behavior such as email, telephone number, hobbies/interests, time spent online, past online purchases	Users most likely to withhold this information [111]

Table 3.3: Type of information users willing/unwilling to share

[107]. An investigation conducted by Willis et al. [192] pursued an analysis of the extent of personalization in Google search results. They utilized "induced" fake profiles by conducting keyword searches and viewing specific YouTube videos. Their underlying expectation was that this information would be used by Google for determination of which ads to display and the reason behind this expectation was Google's policy at the time that stated that ads displayed with search results would be *contextual*

ads, selected only based on information in the search result page itself. However, to their surprise, the searches contained *non-contextual* ads based on inferred interests from previous interactions alongside the contextual ads. Moreover, some of the non-contextual ads could potentially reveal sensitive personal characteristics based on the inferred interests such as an ad containing the question, “Do you have diabetes?” Finally, even when users indicated a preference for personalized search results they did however consider certain queries in their search history to be “sensitive”, and 92% wanted control over what Google was tracking about them as they searched the web [129].

3.1.2 Privacy-Preserving Personalization

Methods that aim to achieve personalization by minimizing potential privacy risks to users are listed and explained in what follows:

- ***Pseudonymous personalization:*** These systems allow users to use pseudonyms in a personalized system. The system can keep track of the pseudonym across different sessions and provide personalized services without accessing the true identity of the pseudonym [17, 80, 94]. The usefulness of pseudonymous personalization is unclear and it remains difficult to achieve when payments, physical goods and/or services’ exchange is involved. Moreover, anonymity of database entries [167], query terms [123], item ratings [65], and textual data [143] can be compromised by a resourceful attacker.

- ***Client-side personalization:*** These systems offer personalization by storage of user data at the client side (e.g. users’ computers or mobile phones) and personalization processing is also undertaken at the client side [48, 68]. On account of user data being located at client side, a variety of certain sophisticated personalization techniques cannot be applied. Moreover, program code that is used for personalization often incorporates confidential business rules/methods, and client-side personalization exposes this confidential information to risk of disclosure.

- ***Distribution, aggregation and obfuscation:*** This class of techniques help protect user privacy in recommendation systems that employ collaborative filtering. The data containing users’ preferences is distributed across users’ own machines, it further updates the user interest model by incorporating one neighbor's ratings at a time and thereafter discards it [114]. Other approaches perform encrypted aggregation of users' data [38] or obfuscate the data through inserting random values into it [26]. The main disadvantage of such methods is with respect to their use being limited to recommendation systems.

- ***User controls and feedback:*** This class of techniques proposes providing users the ability to control what goes into their user model, what information from their model is available to different services, and how the model is managed and maintained [89, 186]. Despite the effectiveness of such controls, their use remains limited on account of the cognitive load imposed by such controls on their users.

3.2 User Modelling

An accurate model for representation of the user is a core element of any personalized application. The underlying process that creates such a representation is termed in the literature as "user modelling [90]." Following lists the three main stages that arise in the context of user modelling [126]:

- *Acquisition* of information about user's interests, preferences, goals, knowledge, behavior and social context so as to ensure "richness" of user data
- *Representation* of the knowledge about the user so as to ensure "creation" of the user profile for use in the personalized application
- *Application* of the user model so as to ensure 'functioning" of the user profile

We discuss works relevant to the above-listed stages in the following subsections.

3.2.1 Acquisition of User-Related Information

This stage of user modelling can be classified in terms of 1) approach used to acquire user-related information, 2) type of user-related information, and 3) scope of user-related information.

The information about user's interests and preferences can be gathered in an implicit or explicit manner. The explicit approach requires user intervention to explicitly communicate to the system his/her interests and preferences. Such intervention generally involves completion of interview forms for specifying interests and preferences

[113], provision of positive or negative relevance feedback about retrieved documents [18, 43], indicating ratings for recommended items [28], or by adding/removing keywords that indicate topical preferences [14]. The implicit approach on the other hand does not require efforts from the user as it gathers information about user's interests and preferences in an unobtrusive manner. This is done through automatic monitoring of the actions undertaken by the user during user-system interaction and typical scenarios involve collection of browser history [24, 105], query history and clickthrough data [66, 157, 160, 163], desktop information [12, 45, 172], document dwell time[1] [47, 197], and user's interactions with social applications [39, 125, 179, 204]. The implicit approach to user modelling attempts to automatically infer user's interests and preferences through the collected user data.

With respect to the type of information stored, there exist two types of information that can be utilized for user modelling, namely user and usage information. User information refers to information about the user such as the user's name, age, language, country, job title, job description and other similar information [23, 156, 110, 187]. Usage information is information that records user behavior in terms of his/her interaction with the system and exists in many forms including queries submitted to the system, clicked results, browsed Web pages, dwell time, items viewed, ratings provided for items etc. [47, 66, 134, 157, 160].

The scope of the acquired user information relates to whether the information gathered is short-term or long-term. Long-term interests represent persistent interests and can be extracted from usage data accumulated over the long run [138, 166, 189, 194]. Inferring these interests from past usage activity and exploiting them can help in satisfying future information needs of a similar nature. As an example, a movie recommendation system can recommend a "new release" action movie to a user who mostly watches action movies while occasionally watching comedy movies, and this can be done based on the user's long-term interest profile. Short-term interests on the other hand are ephemeral interests i.e., those interests that are related to recent activities of the user [24, 161, 180]. As an example, a news recommendation system can infer a user's sudden interest in weather-related news if for example he/she suddenly starts searching for weather-related news information. Few works in the literature capture both long-term and short-term interests of a user [15, 124, 165] in order to facilitate creation of a rich user profile.

[1] Document dwell time is the estimated time that the user spent viewing a document

3.2.2 Representation of User-Related Information

The user model representation concerns aspects related to mechanisms that are used to represent the user's interests. Traditional user models are either vector-based or semantic network-based with terms that are either words/phrases or concepts [67]. Vector-based user models constitute a vector of terms with their associated weights. Semantic network-based models on the other hand consist of nodes and edges with their associated weights that capture terms and their semantically-related or co-occurring terms respectively. Semantic network-based user models provide a richer representation as compared to their vector-based counterparts as they are able to model relationships between terms.

In both models, terms are generally represented by words and/or phrases mined from user or usage information; popular term-weighting schemes include term frequency, term frequency-inverse document frequency and BM25 [20]. In some cases, the terms in the user model can also be represented through conceptual terms which are essentially categorical terms drawn from some knowledge source. The representative knowledge sources could be models developed by domain experts, general knowledge repositories developed by human contributors (such as *Wikipedia*[2]), Web taxonomies (such as *ODP*[3]), or rich ontologies (such as *SUMO*[4]). Another commonly used source is *WordNet*[5] which despite not being a traditional knowledge source is considered as a rich linguistic resource.

The works in [50, 72, 105, 155, 172] mine usage information for extraction of interest terms which correspond to important keywords and these terms are maintained in vectors. Some personalized Web search systems [138, 103, 160] and a more recent contextual advertisement framework [99] in the literature utilize ODP categories and concepts for representation of interest terms in a vector-based model. In recent years the research trend has shifted towards the use of richer knowledge sources that cover a wide range of concepts and their hierarchies e.g. Abel et al. [8] and Orlandi et al. [127] explore aggregation of user data across different Social Web platforms for creation of cross-domain user profiles with semantically enriched concepts extracted from DBPedia. Another very recent work utilizes a Wikipedia-based distributional semantic and entity linking framework to create rich user profiles that are then utilized for context-aware content-based recommendations [121]. On the other end of the

[2] `http://www.wikipedia.org`

[3] Open Directory Project `http://www.dmoz.org`

[4] Suggest Upper Merged Ontology `http://www.ontologyportal.org`

[5] `http://wordnet.princeton.edu`

research spectrum are works that represent user profile data in a semantic network-based model. A well-known example is the system by Micarelli and Sciarrone [113] where stereotypes are associated with a user[6] and these stereotypes are complemented with co-occurring topic terms in documents marked relevant by the user with final merging of active stereotypes (i.e., those that reflect current interests of the user) with passive ones in a graph model. Furthemore, Leung et al. [102] propose the use of concepts (i.e., important terms) that are extracted from the search results of a query, and are organized into a semantic ontology representing the possible topics related to the query; this is followed by employment of user clickthrough data to determine his/her topical preference within the employed ontology.

3.2.3 Application of the User Model

The user model application phase concerns aspects related to means through which a user model is maintained and utilized in the personalized application. To achieve personalization, different approaches for application of the user model have been proposed in the literature and the distinguishing feature between the proposed approaches is the underlying technique used.

One of the earliest approaches utilize a similarity-based technique through the categorization of both user interests and Web search results and a biasing of search results according to some similarity measure on these categories. Approaches along these lines include [46, 103, 104, 105, 169]. Chirita et al. [46] utilized the user model in a graph-based similarity framework over the metadata of ODP category taxonomies in order to perform re-ranking of search results. Liu et al. [103] utilize cosine similarity measure between the ODP categories of the user profile and the ODP categories of the user query. Luxenburger et al. [104] introduced a statistical language model to represent various granularity levels of user search tasks; this language model is then used to compute its similarity with tasks within the same search session and based on this computation the system personalizes adaptively. Matthijs and Radlinski [105] utilize as the user profile a set of features extracted from long-term browsing history after which they apply tf-idf and BM25 based similarity methods on these features. Tan et al. [169] utilized search history language models over long-term user profile data (based on queries, documents and clicks) and re-ranking of search results was performed through the similarity between current query and the corresponding search history model.

[6] Note that these stereotypes are defined by human experts in the domain of Computer Science.

More recently, predictive models have been proposed for application of the user model in personalized applications. State-of-the-art recommendation systems based on collaborative filtering assume an underlying structure to users' rating behavior, and induce predictive models based on past ratings of all users [33]. User-item interactions are modelled with the help of latent characteristics of users and items in the system, models are then trained over the available data and later used to predict ratings of users for new items. Some predictive modelling approaches that are widely utilized in the context of recommendations are Latent Semantic Analysis [82], Latent Dirichlet Allocation [27], Maximum Entropy [207], Support Vector Machines [71], Singular Value Decomposition [96] and more recently neural networks [60]. White et al. [190] constructed a predictive model leveraging short-term contextual information by representing user interests as a list of Open Directory Project (ODP) categories; the strength of the model lies in its ability to scale up to a quarter million users with billions of URLs. Sontag et al. [159] utilized long-term search history-based user profiles in a generative probabilistic model to infer the search relevance of URLs; the model was used to estimate topic-based profiles for both documents and users. Ustinovskiy and Serdyukov [178] proposed a predictive model for query filtering to avoid undesirable harm of personalization on certain queries.

3.2.4 User Modelling Approaches in Lieu of Thesis Contributions

Herein we describe two state-of-the-art user modelling approaches with which we compare our user modelling strategy. We perform this comparison in the context of personalized Web search as explained in detail in Chapter 6. The first approach is the one by Teevan et al. [172] which utilizes a wide variety of user content in an implicit relevance feedback framework for creation of a rich user profile. The second approach is by Matthijs and Radlinski [105] which constructs a user model by means of a user's complete browsing history and refines the model through a parsing method that takes into account web page structure and noun phrases.

Teevan et al. [172] construct a rich user profile by means of the following sources of information:

- A user's previously issued queries
- A user's previously visited web pages
- A user's desktop documents

- A user's created and read emails

The method works by incorporating a relevance feedback mechanism, and instead of explicit relevance judgements, implicit indications of relevance are incorporated by means of creating the user profile from above-listed sources. In other words, Teevan et al. extend traditional relevance feedback via BM25 weighting to incorporate user profile documents as relevance judgements provided by the user. User profile documents containing query terms are considered to be marked as relevant by the user, and the similarity between a query and a document is computed by means of summing query term weights in user profile documents; this similarity is then used to re-rank Web search results.

Matthijs and Radlinski [105] represent a user by means of their long-term browsing history via a list of visited URLs together with the number of visits to each URL, and terms in visited web pages together with their associated weights. The novelty of their approach lies within their intuition to utilize the structure of web pages in a user's browsing history, i.e., they assign higher weights to terms occurring in titles and metadata description of web pages. Moreover, the approach experiments with different term weighting strategies which are then utilized in a re-ranking approach for generating a list of personalized search results for the user. Another significant contribution within the work by Matthijs and Radlinski is an evaluation framework for personalized search that helps evaluation a real-world query workload without putting extra cognitive load on the user.

We utilize the above two approaches as baseline approaches for comparison with our user modelling strategy. Moreover, we make use of the evaluation framework suggested by Matthijs and Radlinski (refer to provided details in Section 2.4.4.2 of Chapter 2).

3.3 Utilization of Social Web Data for User Modelling

The Social Web has revolutionized the Web, transforming it into a dynamic, lively and vibrant medium enabling creation of user-generated content and interaction among Web users. This has opened doors for investigation into new and exciting research challenges e.g. information diffusion [10], community detection [170], influence analysis [40], trend identification and event detection [206] to name a few. According to Paliouras [128], "One of the major innovations in personalization in the last 20 years was the injection of social knowledge into the model of the user", and current research efforts that aim to make use of social Web data for creation of user profiles reinforce this view. The subsections below present an overview of such works.

3.3.1 User Modelling through Social Tagging/Bookmarking Applications

Earlier research efforts that aim to exploit information from online social systems for personalized search rely mostly on social bookmarking and tagging systems [29, 125, 179]. Approaches by Noll and Meinel [125] utilize the notion of frequency of occurrence for tags that users apply to resources (in this case web documents) in order to define a user-document similarity measure that re-ranks the search results. Xu et al. [196] developed a similar personalization method based on user's documents and tags, and also enrich the user profile further through expansion of tags. Vallet et al [179] also utilize a user-document similarity measure based on the *term frequency-inverse document frequency (tf-idf)* scheme in which both the *tf-idf* weights in the user space and document space are calculated for computation of a joint similarity measure. A recent approach by Bouadjenek et al. [29, 30] uses the social bookmarks assigned to documents in a collaborative filtering setting in order to take into account tags used by similar users. Zhou et al. [205] use social tags/bookmarks from within the user profile in a query expansion framework and their method instead of making term-term similarity inferences makes use of an iterative context enhancing and weight propagation mechanism that is able to expand the query with terms belonging to the same topic. Recently, Abel et al. [8] proposed *Mypes* which supports the linkage, aggregation, alignment and semantic enrichment of user profiles available in various Social Web systems, such as Flickr, Delicious and Facebook. These user modelling systems that make use of social tags/bookmarks [8, 29, 125, 179, 196, 205] are less effective for general-purpose search due to the significantly low usage of social bookmarking sites [79] and this is also confirmed by our user-survey based study in chapter 4 [199]. Moreover, these methods do not incorporate rich networking information available within social Web applications.

Another class of social recommendation systems are built upon the principle of *homophily* whereby users' friends on social networks have similar taste [203], and such systems unlike those described in preceding paragraph take into account network information. As an example of such a system, Carmel et al. [39] propose an aggregation tool for information discovery and analysis over the social data gathered from IBM Lotus Connections' applications; documents are retrieved by taking into account social connections of the active user. Moreover, some systems aim to enhance the recommendation process by taking into account preferences indicated by a user's friends within his/her social network [131, 151]. Pera and Ng [131] propose a book recommendation algorithm which combines similarities of tags between candidate books and a target user's favorite books with ratings from friends of the target user. Sharma and Cosley [151] propose a recommendation application (called PopCore) in Facebook that uses several network-centric algorithms based on popularity, similarity and strength with social neighbors. These systems [39, 131, 151] are limited in terms of their usage and applicability to daily information-seeking scenarios on account of the data coming from within an enterprise setting. Twitter on the other hand is publicly accessible for the most part with a powerful API that facilitates academic research. In the following subsection we particularly focus on user modelling techniques utilizing Twitter data which is the focus of this thesis.

3.3.2 Twitter-based User Modelling

Over the past few years, recommendation systems technology has made significant progress and more recently a number of recommendation systems built on top of Twitter data have emerged. Many research directions have been undertaken in this context such as recommendation of hashtags [70, 98], recommendation of URLs [42, 56], news recommendation [52, 132], recommendation of users to follow [73, 76] and tweets' recommendation [58, 62]. From within these directions, we present a brief overview of works closely related to contributions within this thesis.

Abel et al. perform semantic enrichment of tweets in an attempt to model Twitter users [7] for delivery of personalized content in different application domains [5, 4]. Their method however is limited in that the enrichment connects Twitter posts to news articles, and thereby the created user profiles lack coverage.

Nagpal et al. [122] propose the idea of search results' refinement by extracting links (i.e. URLs) from users' "social chatter" whereby "social chatter" refers to users' emails and Twitter feeds. Their method is able to create rich user profiles through exploitation of a target user's email and Twitter network. However, the fact that the personalization process utilizes search indices from URLs of a target user's social chatter renders the user profiles somewhat less useful and incomplete.

Kacem et al. [87] propose a model for dynamic construction of user profiles from Twitter data; their model utilizes a temporal feature that helps serve fresh content when performing personalization. Incorporation of temporal dynamics into the model serves towards creation of a rich user profile that takes into account both long-term and short-term interests. However, their method does not take advantage of the rich networking information available within Twitter.

A very recent technique by Ding and Jian [53] proposes a novel idea that utilizes conditional random fields for the extraction of data from Twitter user biographies and this data is used to reflect users' personal interests. The data extracted from personal user biographies on Twitter despite being accurate provides limited information about the user.

3.4 Personalized Recommendation of Scholarly Works

Over the recent years, there has been an active research interest within the information retrieval community to solve challenges in the domain of digital libraries. A significant research direction in the context of digital libraries is "personal information management and personal digital libraries"; within this domain the generation of personalized recommendations for scholarly works is an emerging challenge and as such taken up as an application domain for the contributions of this thesis. The following subsections discuss related work in the domain of scientific articles' recommendation along with presenting an overview of works that analyze researcher's activities over Twitter.

3.4.1 Scientific Articles' Recommendation

Existing academic search engines (such as Google Scholar, Microsoft Academic Search and CiteSeer) have proven their effectiveness in assisting researchers during the retrieval of scientific articles. However, occasionally these search engines face issues where the retrieved set of articles is either too large or too small [85]. Recommendation systems that generate a reading list of scientific articles have thereby emerged as a popular solution. The most popular approaches either use a set of papers as a query set or a corpus of papers relevant to a given area [171]. One of the earliest works by Woodruff et al. [193] uses a single paper for generation of a reading list through "spreading activation" over its text and citation data. El-Arini and Guestrin [59] use the notion of "influence" to capture the transfer of ideas as individual concepts among papers in the query set. Among the systems that utilize a large corpus of papers to extract core papers of a field, most utilize PageRank over the citation graph along with measures such as "download frequency", "citation count", and "impact factor" [44]. Finally, other techniques utilize collaborative filtering over research papers whereby the user-item ratings' matrix is obtained from the citation network between the papers [108] along with implicit behaviors extracted from a user's access logs [198]. Some works also use a hybrid approach that combines collaborative filtering with content-based filtering methods whereby content analysis is performed through probabilistic topic modeling [11, 184].

For the purposes of our evaluation we utilize a technique proposed by Phelan et al. [132] which utilizes Twitter feeds to recommend to a user news stories from within his subscribed *RSS* feeds. The approach is fundamentally makes use of co-occurring terms from within tweets and news articles and uses the standard *tf-idf* score of those terms across all news articles for the recommendation process. We utilize the approach by Phelan et al. [132] as a standard baseline for comparison with our scientific articles' recommendation framework in Chapter 7.

3.4.2 Analysis of Researchers on Twitter

Recently, researchers have started investigating academic activities on Twitter. The earliest works use scientific tweets as a new measure for citation analysis where citation is defined as a tweet containing a URL to a peer-reviewed scientific article [61, 137, 100, 188]. A more recent work by Hadgu and Jaschke [75] proposes a classification method to construct a directory of computer scientists on Twitter. Their approach starts from a seed set of Twitter accounts from which further Twitter accounts are derived and passed through a machine learning classifier that classifies the Twitter account into researcher or non-researcher. The work by Hadgu and Jaschke [75] can have potentially useful applications from the viewpoint of decreasing the gap between Twitter and science.

3.5 Summary

This chapter covered important related works with regards to the contributions of this thesis. First, we explained various aspects of the "privacy-personalization" paradox while presenting an overview of works that investigated various aspects of users' privacy concerns with respect to personalization; we also presented from within the literature an overview of solutions that aim to provide privacy-preserving personalization. Second, a broad overview of various stages of user modelling pipeline were presented namely acquisition stage, representation stage, and application stage. Third, works related to utilization of Social Web data for user modelling were presented and we also covered Twitter-based user profiles in particular. Fourth, we covered the specific application scenario of providing scholarly recommendations to researchers and within this context works that analyze researchers on Twitter were summarized.

Chapter 4

Predictors of Users' Willingness for Web Search Personalization

Active participation of users on the Social Web enables utilization of this unique form of data for making inferences about users. Such inferences have been successfully deployed in various personalized applications as discussed in Chapter 3 of this thesis. Furthermore, researchers have actively explored the paradox between personalization and privacy. One unexplored area however is the relationship between willingness to use personalized applications and their comfort level with Social Web-based applications. For this thesis, in an attempt to address the first research question raised in section 1.2.1 (Chapter 1), we undertook such an investigation yielding useful insights with respect to users' privacy concerns. It is these insights that were later utilized for the remaining thesis contributions as we explain later in this chapter. We took up personalized Web search as a case-study for an analysis of users' privacy concerns with personalization; the following sections present a detail of our attempt to study the correlations between Web search personalization and users' social network usage patterns. We begin by presenting an overview of the undertaken investigation and its goals in the broader context of pursuing a characterisation of users who prefer personalization. This is followed by an explanation of the survey questions and its various facets; we also present a summary of demographics of the survey respondents along with basic statistics about their online usage (i.e., both with regards to personalization in Web search engines and social network usage). We then present the analyses and findings of the correlation analysis which reveals that Google+ and Twitter serve as significant social networking platforms with regards to providing information about a user's openness to personalization. We also present details of our prediction model that aims to make predictions about a user's willingness for personalization based on features relating to his social network usage. Finally, some limitations and implications of the study are presented before giving the reader two important conclusions driven from this study that form the basis for our Twitter-based user model explained in subsequent chapters.

4.1 Overview

When using a Web search engine to tackle an information-seeking task, users typically oversimplify the complexity of their information needs by selecting a few keywords (query). Moreover, given the lack of information regarding the context of the query, it is hard to attain a reasonable user satisfaction when using Web search engines. In recent years personalized Web search has emerged as a promising way to improve the search quality through customization of search results for people with different information interests and goals. However, concerns about the privacy of users have introduced reluctance in the adoption of personalized Web search systems [93, 185] such as iGoogle [154].

In the pre-digital age the most common way to find useful information was via interaction with friends, colleagues, or domain experts. With the advent of social networking services (SNS) the information-seeking patterns of users have considerably changed [119] leading to an intersection between traditional and modern approaches of information-seeking [118]. Social networking services are now rekindling the pre-digital information-seeking pattern in the digital world.

Over the past few years, many research efforts have focused on both personalized Web search and social search [57, 84]. We investigate the correlation between social network usage patterns of users and their openness to opt for Web search personalization. This can open doors for understanding the user's willingness to adopt Web search personalization based on observable correlations. Re-introducing the motivating example from Chapter 1, consider the following two scenarios:

- User A is highly active on various social networking services as he/she communicates his/her thoughts over a range of topics (politics, religion, economics etc.)

- User B is much less active when compared to user A in sharing his/her thoughts on social networking services.

In line with this scenario it would be interesting to investigate the correlations of the behaviors of both users A and B and their openness to Web search personalization. Some commercial systems have attempted a similar integration (as is evident in recent social search approaches by Google and Bing: e.g., Bing's Facebook integration[1] and Google's "Search Plus Your World" Google+ integration[2]). However, to the best of our knowledge, there is no literature that describes the correlations that we investigate in this thesis.

Apparently, the existence of a correlation between the willingness to personalize Web search and the inclination to use social networks seems intuitively obvious and hence, trivial. However, a thorough analysis of the various usage patterns within different types of popular social networking services is something that needs careful investigation so as to form a more coherent basis for the development of meaningful and well-accepted personalized search systems. With the aim of finding a characterisation of users who would prefer personalized search results more than non-personalized search results we conducted a user survey. We designed the user survey so that we can investigate the social network usage patterns of users together with their privacy concerns with respect to Web search personalization, and their preferences to opt for personalized Web search. We also looked into various SNS tools (more specifically Facebook, Twitter, LinkedIn etc.) as well as at the characteristics of SNS usage (such as frequency of SNS usage, frequency of posting SNS updates, number of friends on SNS, frequency of asking questions on SNS). We hypothesise these features might relate more closely to users' willingness for Web search personalization. We conducted a large-scale user survey in two parts where the first part gathered responses from 380 people from various countries, and the second part gathered responses from 113 people from various countries. This data was then used in a regression model to analyse the correlation between SNS variables and willingness to personalize Web search. The data was also used in a support vector machine (SVM) model to explore the potential to make predictions about users who would be willing to opt for personalized Web search results. We wish to explore if the prediction accuracy would be sufficient for a real personalised Web search system. In these analyses we discovered a number of useful patterns and behaviours about users' openness to Web search personalization; these outcomes can be of help in the future developments of personalized Web search and social search systems.

[1] http://www.bing.com/community/site_blogs/b/search/archive/2011/05/16/news-announcement-may-17.aspx

[2] http://www.google.com/insidesearch/plus.html

4.2 Survey and Survey Results

In this section we describe the survey methodology and the measures and variables that are analysed. The survey comprised 20 close-ended questions with five questions of a general nature (collecting basic information about the participants), five questions related to various aspects of Web search personalization (explained in section 4.2.2) and ten questions related to SNS usage (explained in section 4.2.2). The entire survey is listed in Appendix A of this thesis.

4.2.1 Participants and Survey Content

In order to understand how SNS usage patterns affect people's willingness to personalize Web search results, we designed a survey and dispatched it to a wide range of people in various countries (i.e. Ireland, Italy, Spain, France, United Kingdom, Finland, United States, Canada, Pakistan, India, South Korea). In the first phase, the survey was completed by 380 people. Demographic characteristics for the survey respondents in the first phase are shown in Table 4.1. Participants were recruited via university distribution lists (both online and offline) and social networking sites (chiefly, Facebook and Twitter) and we recruited a diverse range of people; distribution lists were employed to avoid recruiting only those participants who have a social network presence and thereby avoiding high skew in the results. In addition to collecting basic demographic information, we also collected information about participants' use of SNS tools such as which SNS accounts they have and which of them they use the most (shown in Table 4.2 and Table 4.3 respectively).

Demographics	N (%)
Male	235 (61.8%)
Female	145 (38.2%)
Europe	206 (54.2%)
America	21 (5.5%)
Asia	153 (40.3%)
10-20	0 (0%)
21-30	259 (68.2%)
31-40	87 (22.9%)
41-50	19 (5%)
Above 50	15 (3.9%)

Table 4.1: Demographic Variables (n=380)

SNS Tool Details	N (%)
Facebook Presence	356 (93.7%)
Twitter Presence	241 (63.4%)
Google+ Presence	239 (62.9%)
LinkedIn Presence	272 (71.6%)
Bookmarking Sites Presence	60 (15.8%)

Table 4.2: Statistics for SNS Accounts of Survey Respondents

SNS Usage Details	N (%)
Facebook As Most Used	325 (85.5%)
Twitter As Most Used	106 (27.9%)
Google+ As Most Used	30 (7.9%)
LinkedIn As Most Used	17 (4.5%)

Table 4.3: Statistics for Highly Used SNS Accounts by Survey Respondents

4.2.2 Measures and Variables

The following variables were included in the analyses:

4.2.2.1 Personalization Response

Respondents were asked whether or not they considered personalized search results to be of any benefit to them[3]. This was a binary variable with a "Yes" or "No" response. Additionally, a likert-scale variable was used corresponding to respondents' agreement with Web search personalization making the information-seeking process less painstaking; the scale ranged from "Strongly Disagree" (1) to "Strongly Agree" (5). Furthermore, three additional binary variables related to Web search personalization features of current search engines were also investigated. Respondents were asked about their awareness of the personalization feature in existing Web search engines in addition to their awareness about search engines making use of their search history data for the process of Web search personalization and finally, whether or not they were comfortable with such use. We report these statistics about Web search personalization in Table 4.4.

4.2.2.2 Facebook Usage

Considering the high usage of Facebook as reported in Table 4.3, some questions in the survey were particularly focused towards Facebook usage. In particular, respondents were asked about the frequency of their Facebook usage (with scale ranging from "several times a day" (7) to "never" (1)), frequency of posting something (status update, photo or link) on Facebook (with scale ranging from "frequently" (4) to "never" (1)), frequency of liking something on Facebook (with scale ranging from

[3] For the ease of survey respondents, we included in the survey an explanation of what Web search personalization is along with an explanation of how implicit user data is used for this process. Furthermore, we asked the survey respondents to contact us in case of any confusion with respect to Web search personalization and some of them asked us questions to understand the personalization process better.

Personalized Web Search Results Considered as Useful	N (%)
Yes	188 (49.5%)
No	192 (50.5%)
Awareness about Personalization Feature of Web Search Engines	N (%)
Yes	275 (72.4%)
No	105 (27.6%)
Awareness about Web Search Engines using Search History Data	N (%)
Yes	337 (88.7%)
No	43 (11.3%)
Comfortable with Web Search Engines using Search History Data	N (%)
Yes	196 (51.6%)
No	184 (48.4%)
Likert Scale for Agreement on Worth of Web Search Personalization	MEAN (SD)
	3.58 (0.98)

Table 4.4: Statistics on Users' Desirability for Web Search Personalization

"frequently" (4) to "never" (1)) and the approximate number of Facebook friends. We report these statistics about Facebook usage in Table 4.5.

4.2.2.3 Twitter Usage

We also included some Twitter-specific measures in our analysis. The survey did not ask for these measures explicitly. Instead, the survey respondents who used Twitter were asked to provide their Twitter handles which were then used to fetch all their tweets. From these tweets, we extracted for the survey respondents their number of mentions (the mention feature of Twitter enables its users to address a specific user within a tweet) and number of retweets (the retweet feature of Twitter enables its users to re-post a tweet posted by someone else). Additionally we also extracted the number of topics contained in survey respondents' tweets through the use of Twitter-LDA [202].

Frequency of Posting on Facebook	MEAN (SD)
	2.99 (0.95)
Frequency of Facebook Likes	MEAN (SD)
	3.22 (0.94)
No. of Facebook Friends	N (%)
Less than 100	62 (17.4%)
100-200	88 (24.7%)
200-300	85 (23.9%)
300-400	50 (14.0%)
400-500	28 (7.9%)
More than 500	43 (12.1%)

Table 4.5: Statistics on Facebook Usage

4.2.2.4 Q & A Activity on SNS Tools

Lastly, given the significance of Q & A activity on SNS [119], the survey also included questions about users' Q & A activities on SNS tools. Respondents were asked whether they had ever used SNS tools for information-seeking and whether they considered Q & A activity on SNS as useful. These were binary variables with a "Yes" or "No" response. If respondents preferred Q & A activity on SNS, we further asked them about their frequency of asking questions on SNS along with the frequency with which they considered answers coming from SNS as more reliable than answers obtained from search engines (with scale ranging from "most of the time" (4) to "never" (1)). We report these statistics about Q & A activity on SNS in Table 4.6.

Ever Used Social Networks for Information-Seeking	N (%)
Yes	272 (71.6%)
No	108 (28.4%)
Q & A Activity on Social Networks Considered as Useful	N (%)
Yes	187 (49.2%)
No	193 (50.8%)
Frequency of Asking Questions on SNS	MEAN (SD)
	1.99 (0.95)
Frequency of Considering Answers on SNS More Reliable than Search Engines	MEAN (SD)
	2.38 (1.03)

Table 4.6: Statistics on Social Network Q & A Activity

4.3 Analyses and Findings

We first examine the associations between variables representing willingness towards Web search personalization while controlling for demographic and other factors. Table 4.7 shows predictors of acts for such willingness where the rows of the table represent these acts. Here, the acts are basically the variables corresponding to presence on various SNS tools, frequency of usage of various SNS tools, Facebook usage, Twitter usage, Q & A Activity on SNS as explained in section 4.2.2. Table 4.7 shows the results of logistic regressions with binary outcomes in the following dependent variables:

- User's trust in search personalization as a beneficial process (shown as *WP i.e., Willingness of Personalization* in Table 4.7)
- User's awareness of personalized search services such as iGoogle (shown as *AP i.e., Awareness of Personalization* in Table 4.7)
- User's awareness that search engines use their search history data for the process of Web search personalization (shown as *AH i.e., Awareness of History* in Table 4.7)

- User's acceptance (comfort level) of the fact that search engines use their search history data for the process of Web search personalization (shown as *WH i.e., Willingness of History* in Table 4.7)

1		WP	AP	AH	WH
2	Male	1.635**	2.643**	6.318***	1.480*
3	Female	0.0006**	0.378**	0.158***	0.676*
4	American	0.001	0.004	4.478	0.002
5	Asian	0.001	0.003	0.004	0.001
6	European	0.001	0.004	0.001	0.001
7	Age	1.069	0.997	0.981	0.841
8	Facebook Presence	0.982	0.561	0.860	1.844
9	Twitter Presence	1.544*	1.692	3.651***	1.519*
10	Google+ Presence	1.816***	1.427	1.364	1.626**
11	LinkedIn Presence	0.940	2.475*	2.031**	1.150
12	Bookmarking Sites Presence	1.289	2.535	1.153	1.771**
13	High Usage of Facebook	1.599	0.736	0.488	1.222
14	High Usage of Twitter	1.166*	1.574	3.986**	1.353
15	High Usage of Google+	3.042***	1.565	0.637	2.292**
16	High Usage of LinkedIn	1.193	2.069	1.127	0.971
17	Facebook Usage Frequency	0.898	1.204	1.051	1.231
18	Facebook Posting Frequency	1.637***	0.450*	0.893	1.246
19	Facebook Liking Frequency	0.920	1.776	0.922	0.924
20	No. of Facebook Friends	0.873*	1.031	1.181	0.899
21	Twitter Mentions	1.000	0.983	0.997	0.998*
22	Twitter Retweets	1.001	0.982	0.999	0.999
23	No. of Topics in Tweets	0.997	0.969	1.036	1.012
24	No. of Tweets	0.999	1.015*	1.001	1.001*
25	Prefers Q & A Activity on SNS	1.821***	1.474	0.972	1.492*
26	Considers Q & A Activity on SNS as Useful	1.771***	1.173	0.916	1.318
27	Frequency of Q & A Activity on SNS	1.366**	0.940	0.884	1.273**
28	Frequency of Considering Responses				
29	from SNS More Useful than Search Engines	1.374***	1.232	0.950	1.228**

Note *p<.05, **p<.01, ***p<.001

Table 4.7: Logit regression showing the odds of users' willingness towards Web search personalization

The results from Table 4.7 show that males are more likely to consider Web search personalization as beneficial (row 2 corresponding to *WP*) while location and age do not have much effect on willingness for Web search personalization (row 3-6 corresponding to *WP*). Furthermore, the presence of a user on Twitter and/or Google+ is a strong indicator that he/she will consider Web search personalization as beneficial and similar is the case for his/her high usage of Twitter and/or Google+ (row 8-9 corresponding to *WP* and row 13-14 corresponding to *WP*); the increase is more significant for user presence on Google+ and for high usage of Google+. Additionally, as expected an increase in posting frequency on Facebook increases the odds of willingness to personalize Web search and contrary to expectation, an increase in the number of users' friends on Facebook decreases the odds of willingness

	Personalization Agreement
High Usage of Facebook	0.278*
High Usage of Twitter	0.143
High Usage of Google+	0.338*
High Usage of LinkedIn	0.036
Facebook Usage Frequency	0.052
Facebook Posting Frequency	0.139***
Facebook Liking Frequency	0.108*
No. of Facebook Friends	-0.026
Twitter Mentions	-0.171*
Twitter Retweets	-0.188
No. of Topics in Tweets	0.148
No. of Tweets	0.145*
Prefers Q & A Activity on SNS	0.051
Considers Q & A Activity on SNS as Useful	0.083
Frequency of Q & A Activity on SNS	0.165**
Frequency of Considering Responses from SNS More Useful than Search Engines	0.025

Note *p<.05, **p<.01, ***p<.001

Table 4.8: OLS regression showing the level of users' agreement towards Web search personalization

to personalize Web search (row 17 and row 19 corresponding to *WP*). Lastly, users who use SNS more frequently for Q & A activities are more likely to be willing to opt for Web search personalization in addition to users who frequently consider that responses coming from SNS as more reliable than responses from search engines (row 24-27 corresponding to *WP*).

For the dependent variable reflecting whether or not the user is aware of the personalization feature in current Web search engines (AP), the likelihood is increased if the user is a male, has a presence on LinkedIn, has a high posting frequency on Facebook, and has a high number of tweets on Twitter. Similarly for the dependent variable reflecting whether or not the user is aware of search engines making use of his/her search history data for the process of Web search personalization (AH), the likelihood is increased for males, for users with Twitter and/or LinkedIn presence, and for highly frequent Twitter users. Lastly, for the dependent variable reflecting whether or not the user is comfortable with search engines making use of his/her search history data for the process of Web search personalization *WH*, the likelihood is increased for males, for users of Twitter, Google+ and/or bookmarking sites, for more frequent Google+ users, for users with less mentions and/or high number of tweets on Twitter, for users who frequently use SNS for Q & A activities and consider responses from SNS to be more reliable than responses coming from search engines.

Table 4.8 shows that users having a high usage frequency on Facebook and/or Google+ tend to agree more strongly with the notion of Web search personalization making the information-seeking process easier and less painstaking. This is also the case for users who have a high posting and liking frequency on Facebook along with a high tweeting frequency on Twitter. On the other hand, users' level of agreement with the notion that Web search personalization makes the information-seeking process easier decreases with their frequency of mentions on Twitter. Lastly, users with a high frequency of Q & A activity on SNS agree to a higher degree with the notion of Web search personalization making the information-seeking process easier.

4.4 Prediction Model for Web Search Personalization Willingness

In this section we explain the prediction model that we developed to predict whether or not a user would be interested in Web search personalization. We performed the predictions on a second set of user survey data (test data for our prediction model) and for this we collected responses of 113 people using the same survey designed for the correlation analysis of section 4.2 and 4.3. The data collected in the first phase of the survey served as the training data (i.e., the 380 responses gathered for correlation analysis) for our model.

We utilized five types of information described in section 4.2 that can help characterize a user's willingness for Web search personalization. Here, we refer to the demographic features (Table 4.1) as *demographics*, features denoting presence on various SNS tools (Table 4.2) as *sns_presence*, features describing high usage of various SNS tools (Table 4.3) as *sns_highusage*, features describing Facebook usage (Table 4.5) as *fb_usage* and features describing Q & A activity (Table 4.6) as *qa_activity*. The features we explored are used in conjunction with a supervised machine learning framework providing a model for predicting users who would (and wouldn't) opt for personalized Web search results. The baseline approach we use is to set all Web search personalization predictions to "having no willingess for Web search personalization." As a learning algorithm, we used Support Vector Machines which is a state-of-the-art machine learning algorithm; we utilize SVM with linear kernel which is the simplest setting. We found the set of features that best predicted the variable *WP i.e., Willingness of Personalization* (shown in Table 4.7). The goal of our prediction model is two-fold: first, to see if the prediction accuracy would be sufficient for a real personalized Web search system and second, to explore the value of various types of information in the process of automatically determining willingness for Web search personalization.

Figure 4.1 shows the classification performance of the personalization willingness prediction model, incrementally examining the role of each feature in the prediction accuracy[4]. In-line with intuition, we witnessed a consistent, gradual increase in performance as additional information is made available to the classifier. It is interesting to note the significant performance boost when we move from the baseline to *sns_presence* and *fb_usage* and also when the *qa_activity* features are added to the model. We first examined each feature in isolation (a maximum of 53.9% was obtained with *sns_presence*); after this we applied all possible permutations of features and Figure 4.1 shows the best possible order. Furthermore, within each category the following features led to the biggest performance gain:

- *demographics*: Gender, Location and Age
- *sns_presence*: Twitter Presence, Google+ Presence and Bookmarking Sites Presence
- *sns_highusage*: High Usage of Google+
- *fb_usage*: Facebook Usage Frequency, Facebook Posting Frequency and No. of Facebook Friends
- *qa_activity*: Prefers Q & A Activity on SNS, Considers Q & A Activity on SNS as Useful, Frequency of Q & A Activity on SNS and Frequency of Considering Responses from SNS More Useful than Search Engines

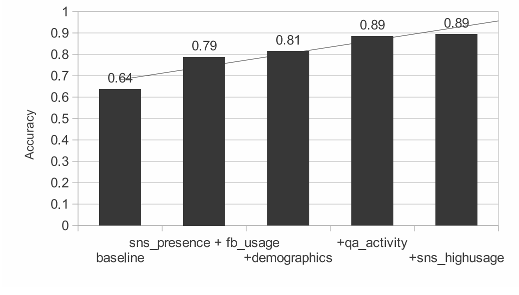


Figure 4.1: Overall classification results for Various Feature Types

[4] We show the combinations that are statistically significant at the 0.95 level with respect to the baseline.

4.5 Discussion

We will now discuss the theoretical and practical implications of our study together with some limitations. Our data show that the percentage of survey respondents who do not consider Web search personalization as beneficial is higher than expected; similar is the case for survey respondents who are not comfortable with Web search engines using their search history data (refer to Table 4.4 for these statistics). The presence of such privacy concerns highlights the need for the undertaken investigation. The correlations (and their results) investigated here may seem intuitively obvious in terms of high engagement in social networks signifying a higher degree of readiness to accept the (at least partial) loss of privacy that is inevitably involved in search personalization. However, recent research indicates that this is not the case as users of SNS have shown growing privacy concerns [164]. Boyd and Hargittai [32] found that the majority of young adult users of Facebook have engaged with managing their privacy settings on the site at least to some extent and noted a rise in such privacy settings' engagement between 2009 and 2010, a year in which Facebook unveiled many controversial privacy changes that made more of information on the site public.

4.5.1 Limitations

Despite our efforts to recruit a diverse range of participants for the study the results may be somewhat skewed towards considerable levels of use of SNS (Table 4.2) and this may not be representative of the general population. Another possible limitation may be the users' lack of information about their answers to the survey questions due to not having sufficient insight regarding what personalized search actually comprises [93]. Nevertheless, there are reasons to be confident in our conclusions. First, the high skewness towards SNS is natural given the immense popularity of these services and evidence suggests that almost every Web user maintains SNS presence (the important difference lies in SNS activity which is a variable thoroughly investigated by us). Second, as explained previously we thoroughly explained the process of Web search personalization to the users so as to inform them of privacy considerations that arise from it.

4.5.2 Implications

Users' privacy concerns have proven to be a significant challenge with respect to Web search personalization, and the issue of when to personalize and when not personalize represents an important challenge, which has not been deeply investigated yet.

Preliminary investigations by Teevan et al. [173] have demonstrated personalization to be topic dependent whereby this dependency arises from variations in user intent. We however argue for a more user dependent notion of personalization on account of the tendency of different users to exhibit varying levels of privacy concerns as has been shown through research on social networks' privacy analysis [93, 164, 176]. As an example, a certain user may not appreciate the fact that a previously visited medical Web site is shown in response to his medical query[5] while another user may be perfectly fine with search results containing previously accessed Web sites.

We argue through the investigations in this chapter that social network usage patterns of users can serve as a significant predictor for determination when to personalize and when not to personalize. Understanding the target audience of personalized search systems is an important aspect for the development of meaningful and well-accepted systems, and hence, this work serves as a first step in that dimension. This could aid towards removing the need for the user to state privacy requirements and to infer these settings through social network usage patterns.

The decision to investigate the correlation between social network usage patterns and willingness for Web search personalization was taken on account of many studies that have shown people to spend a considerable amount of time on social networks [54]. This high amount of social network usage by users in turn has also affected their information-seeking habits and specifically, the way they interact with search engines [117]. Hence, we argue that it is necessary to revisit the notion of Web search personalization from this perspective.

4.6 Conclusion

From the point of view of contributions in this thesis, the following significant conclusions that were drawn from the investigation conducted in this chapter were utilized to develop our user modelling technique:

- Social bookmarking sites are no longer in widespread use as is evident from Table 4.2 with a very low presence of 15.8% among the participants. Hence, one can easily conclude that personalized solutions built on top of social bookmarks/tags will be less effective.

[5] The user may feel as if his privacy has been violated through search results' monitoring, and such violation may be a highly sensitive matter for the user on account of the query involving his/her medical history.

- Twitter and Google+ serve as important indicators for personalization willingness as Table 4.7 shows. Presence and high usage of these two social networking platforms indicate willingness to adapt personalization; and as an outcome of this particular finding we utilize Twitter for our user modelling framework.

4.7 Summary

In this chapter, we utilized a survey methodology to gather relevant data and then investigated the correlations between users' social network usage patterns and their openness to opt for Web search personalization. The participants' responses to the survey questions enabled us to use a regression model for identifying the relationship between SNS variables and willingness to personalize Web search. We also performed a follow-up user survey for use in a support vector machine (SVM) based prediction framework. The prediction results lead to the observation that SNS features such as a user's demographic factors (such as age, gender, location), a user's presence or absence on Twitter and Google+, amount of activity on Twitter and Google+ along with the user's tendency to ask questions on social networks are significant predictors in characterising users who would be willing to opt for personalized Web search results. By exploiting SNS data, predictions can be made about user preferences without requiring explicit user input.

Chapter 5

Modelling User Interests through Twitter

In this chapter the usefulness of user profiling based on users' generated contents and users' behaviors in Twitter is considered. In particular, we analyse the language and the behavior of users in their microblog generation to define a user model. In an attempt to address the second and third research question raised in section 1.2.1 (Chapter 1), a language modelling approach to the creation of user profiles is presented, which leverages both the content that users generate in Twitter, and the content generated by other users to whom these users are connected in the social network. The chapter begins with an overview of the model while also presenting its novel aspects. This is followed by details of the user modelling strategy within which we first explain various Twitter-specific behaviors that the model is based upon along with the individual components of the model (namely, user's own tweets together with tweets of those user mentions, retweets and follows). Finally, similarity measures that depict the degree of similarity between the connected users and the user being modelled are explained in detail.

5.1 Model Overview

We first present an overview of the proposed user model along the various dimensions introduced in section 3.2 of Chapter 3. Note that we utilize the notation introduced in section 2.1 of Chapter 2 to explain various aspects of the proposed user modelling strategy.

The proposed model gathers user-related information in an implicit manner without requiring any intervention from the user. The user-related information acquired by the model falls into the category of usage information in that the model collects user interaction with the system. More specifically, the model records microblog activity by a user which includes content (tweets) he/she generates in addition to content (tweets) generated by those users whom he/she follows, mentions, and retweets. We model a user's long-term interests through the acquisition of tweets over a period of time; however, the model is rich in this particular aspect can be controlled by decreasing the window of time for the tweets under consideration. The user model representation constitutes a vector of terms (from within the collected tweets). From an application perspective, the proposed model is similarity-based as it re-ranks based on language model-based similarity between user profile data and returned documents.

5.1.1 Novelty

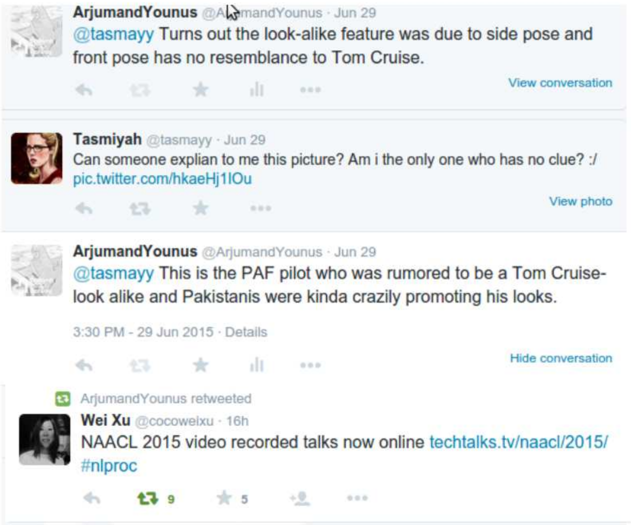
Microblogs such as Twitter constitute an alternate and effective form of user-generated content while also serving as a platform for users' social networking activities. To the best of our knowledge, the use of microblogging platforms and, in particular, Twitter has not, with the exception of a few works [7, 53, 87, 200], been explored as a source of user profile construction for personalized applications and we undertake such a direction in this work. In particular, the language employed by the user and the language of users who are trusted by the user offer a basis for modelling the user's interests and preferences: Twitter provides access to both these sources of evidence which we use for creation of a rich user profile.

As explained in Chapter 3, traditional sources of user modelling evidence comprise log data such as search query logs and clickthrough data, items viewed and rated, document dwell time etc.; we however demonstrate the feasibility of user profiles that model the user's language and related interests through utilization of their microblog (i.e., Twitter) generation and usage patterns. In particular, aspects of a user's Twitter network (users followed, mentioned, retweeted) are used in a language modelling framework to generate a user model while also incorporating similarity measures between a user and his/her network based upon user interactions on Twitter.

The novel aspects of our proposed user modelling strategy are listed as follows:

- We propose a statistical language modelling approach for user profile construction which takes into account various features of a target user's Twitter network and his behavior on Twitter.

- We enhance the model through application of a binary inclusion factor incorporated into the model. This factor (as we explain later in this chapter) introduces



ArjumandYounus @ArjumandYounus · Jun 29
@tasmayy Turns out the look-alike feature was due to side pose and front pose has no resemblance to Tom Cruise.

Tasmiyah @tasmayy · Jun 29
Can someone explian to me this picture? Am i the only one who has no clue? :/
pic.twitter.com/hkaeHj1lOu

ArjumandYounus @ArjumandYounus · Jun 29
@tasmayy This is the PAF pilot who was rumored to be a Tom Cruise-look alike and Pakistanis were kinda crazily promoting his looks.

3:30 PM - 29 Jun 2015 · Details

ArjumandYounus retweeted
Wei Xu @cocoweixu · 16h
NAACL 2015 video recorded talks now online techtalks.tv/naacl/2015/
#nlproc

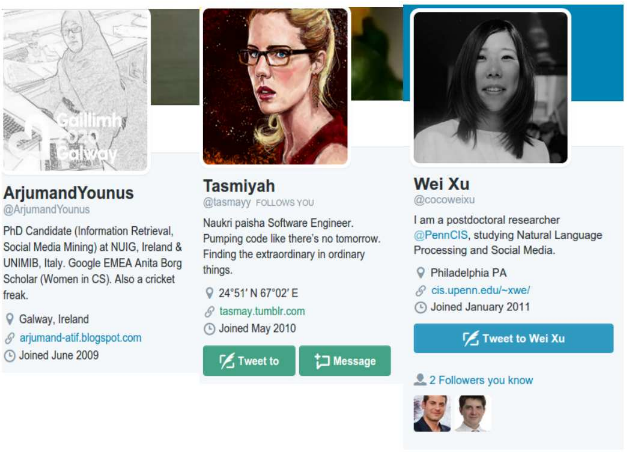
Figure 5.1: Mention and Retweet by Target User

into the model the ability to either allow binary inclusion of Twitter users connected to the user being modelled or incorporation of a similarity-based weight as an indication of how similar connected Twitter users and the user being modelled are.

- We propose different weighting strategies within the language model based on a user's similarity with his/her Twitter network. We incorporate two versions of similarity measures into the model, namely network-based similarity and topical similarity.

5.2 Details of Twitter-based User Modelling Strategy

This section presents in detail the proposed user profiling model that we build using a user's Twitter network. We first present a brief overview of the Twitter-specific behaviors followed by a presentation of the formulations for incorporation of each of the individual Twitter models. Finally, we propose two classes of similarity measures as further extensions to the proposed model.



ArjumandYounus
@ArjumandYounus
PhD Candidate (Information Retrieval, Social Media Mining) at NUIG, Ireland & UNIMIB, Italy. Google EMEA Anita Borg Scholar (Women in CS). Also a cricket freak.
Galway, Ireland
arjumand-atif.blogspot.com
Joined June 2009

Tasmiyah
@tasmayy FOLLOWS YOU
Naukri paisha Software Engineer. Pumping code like there's no tomorrow. Finding the extraordinary in ordinary things.
24°51′ N 67°02′ E
tasmay.tumblr.com
Joined May 2010

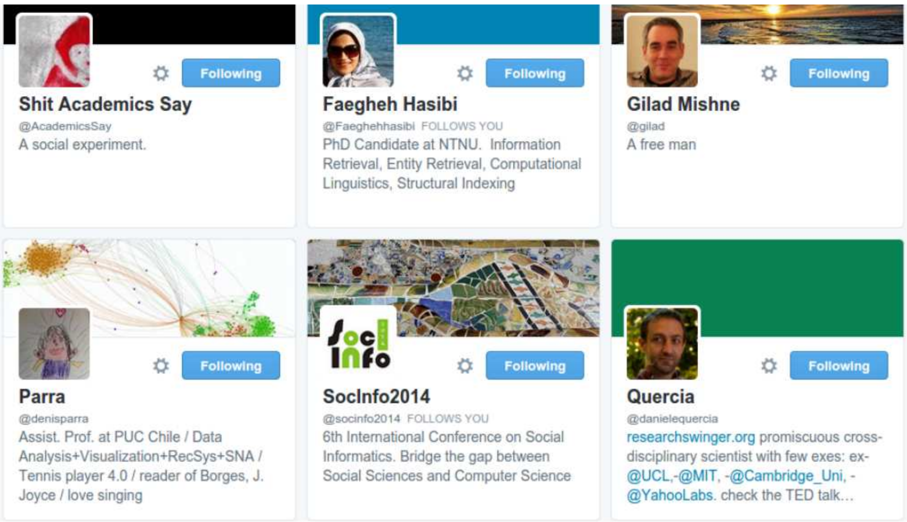
Wei Xu
@cocoweixu
I am a postdoctoral researcher @PennCIS, studying Natural Language Processing and Social Media.
Philadelphia PA
cis.upenn.edu/~xwe/
Joined January 2011

Figure 5.2: Twitter Biography of Target User, User Mentioned by Target User and User Retweeted by Target User (from left to right)

5.2.1 Twitter-specific behaviors

As a recap from Chapter 2, we remind the reader that the Twitter microblog network enables a user to follow any other user and unlike most online social networking sites the relationships of *following* and *being followed* require no reciprocation. Being a follower on Twitter enables a user to receive all the messages (called tweets) from those members the user follows. Twitter presents the opportunity to users to post 140-character long status updates about a variety of topics. Twitter also enables users to engage in conversations with each other through a feature known as *mentions* while at the same time allowing users to share a tweet written by another Twitter user with his/her followers through a feature known as *retweets*.

The above Twitter-specific behaviors form the core of our approach as they aid the incorporation of the language of trustworthy users in order to identify interests and preferences of the user under consideration. From this point onwards in the thesis we use the phrase "target user" to refer to the user under consideration whose profile we wish to create. We incorporate the *mention*, *retweet* and *follow* features of Twitter within our model with the underlying intuition that those Twitterers the target user mentions, retweets or follows may reflect, to a large extent, the target user's own preferences and interests. In order to aid the reader's understanding of the significance of users followed, mentioned, and retweeted we present few examples in Figures 5.1, 5.2, and 5.3. The target user in this example is the Twitter user with screen name "ArjumandYounus" i.e. the author of this thesis. Figure 5.1 shows a mention activity and a retweet activity by the target user; the target user mentions the user with



Shit Academics Say
@AcademicsSay
A social experiment.

Faegheh Hasibi
@Faeghehhasibi FOLLOWS YOU
PhD Candidate at NTNU. Information Retrieval, Entity Retrieval, Computational Linguistics, Structural Indexing

Gilad Mishne
@gilad
A free man

Parra
@denisparra
Assist. Prof. at PUC Chile / Data Analysis+Visualization+RecSys+SNA / Tennis player 4.0 / reader of Borges, J. Joyce / love singing

SocInfo2014
@socinfo2014 FOLLOWS YOU
6th International Conference on Social Informatics. Bridge the gap between Social Sciences and Computer Science

Quercia
@danielequercia
researchswinger.org promiscuous cross-disciplinary scientist with few exes: ex-@UCL,-@MIT, -@Cambridge_Uni, -@YahooLabs. check the TED talk...

Figure 5.3: Users Followed by the Target User

screen name "tasmayy" and retweets the user with screen name "cocoweixu". Figure 5.2 shows the Twitter biographies of the target user along with the user she mentions (i.e., "tasmayy") and user she retweets (i.e., "cocoweixu"); from these biographies it is evident that the user she mentions reflects the interests of the target user due to belonging to same hometown (i.e., city of Karachi with latitude and longitude of 24°51′ N and 67°2′ E respectively.) while the user she retweets reflects her interests on account of overlap in a research field (i.e., social media). Figure 5.3 shows a small sample of six users[1] followed by the target user and as can be seen four of the six followed users (i.e., users with screen name "Faeghehhasibi", "gilad", "denisparra", and "danielequercia") are academics working in the same domain as the target user, one is a comic account related to things academics say with respect to their research life (i.e., user with screen name "AcademicsSay"), and another is a Twitter account for a scientific conference (i.e., user with screen name "socinfo2014"). All these followed users reflect the interests and preferences of the target user.

5.2.2 Individual Twitter models that constitute the overall model

The proposed model utilizes the tweets of the target user along with the tweets of those users whom the target user mentions, retweets or follows.

[1] Twitter allows a particular user to follow as many users as he wishes; however, due to space limitations we show only six of these.

Equations 5.1, 5.2, 5.3, and 5.4 below show how we include into the model content generated by 1) the target user, 2) users whom the target user mentions, 3) users whom the target user retweets, and 4) users whom the target user follows.

$$P(w \mid T_o) = \frac{1}{|T_o|} \sum_{t \in T_o} P(w \mid t) \tag{5.1}$$

$$P(w \mid T_{U_m}) = \frac{1}{|U_m|} \sum_{u_i \in U_m} \frac{sim(u, u_i)}{|T_{u_i}|} \sum_{t \in T_{u_i}} P(w \mid t) \tag{5.2}$$

$$P(w \mid T_{U_r}) = \frac{1}{|U_r|} \sum_{u_i \in U_r} \frac{sim(u, u_i)}{|T_{u_i}|} \sum_{t \in T_{u_i}} P(w \mid t) \tag{5.3}$$

$$P(w \mid T_{U_f}) = \frac{1}{|U_f|} \sum_{u_i \in U_f} \frac{sim(u, u_i)}{|T_{u_i}|} \sum_{t \in T_{u_i}} P(w \mid t) \tag{5.4}$$

Here, in case of Equation 5.1 w denotes a word under consideration from within the vocabulary and T$_o$ denotes the set of original tweets by the target user u. In case of Equation 5.2, U_m denotes the set of users whom the target user mentions, T$_{U_m}$ denotes the tweets by those Twitterers whom the target user u mentions (i.e., Twitterers in set U_m), and *sim(u,u$_i$)* denotes the similarity between the target user u for whom we want to create the user profile and each user u_i occurring in U_m. In case of Equation 5.3, U_r denotes the set of users whom the target user retweets, T$_{U_r}$ denotes the tweets by those Twitterers whom the target user u retweets (i.e., Twitterers in set U_r), *sim(u,u$_i$)* denotes the similarity between the target user u for whom we want to create the user profile and each user u_i occurring in U_r. In case of Equation 5.4, U_f denotes the set of users whom the target user follows, T$_{U_f}$ denotes the tweets by those Twitterers whom the target user u follows (i.e., Twitterers in set U_f), *sim(u,u$_i$)* denotes the similarity between the target user u for whom we want to create the user profile and each user u_i occurring in U_f. The computation for similarities *sim(u,u$_i$)* in each of the above equations is detailed in the next Section. Finally, in each of these Equations *P(w | t)* denotes the probability of w in tweets t using Dirichlet prior smoothing:

$$P(w \mid t) = \frac{n(w,t) + \mu \frac{n(w, coll)}{|coll|}}{|t| + \mu}$$

where *n(w,.)* denotes the frequency of word w in (.), *coll* is short for collection which refers to all tweets by the target user u (in case of equation 5.1), all tweets by Twitterers in set U_m (in case of equation 5.2), all tweets by Twitterers in set U_r (in case of equation 5.3), and all tweets by Twitterers in set U_f (in case of equation 5.4), $|.|$ is the overall length of the tweet or the collection, and μ is the Dirichlet prior which we set to the standard value of '2000'.

The individual Twitter models constituting the overall model are summarized as follows:

- Equation 5.1 enables us to model the probability distribution of words within the target user's language that he/she employs over Twitter and, as shown by [6], this language captures the interests of a user.

- Equation 5.2 provides a model for the probability distribution of words used by those Twitterers whom the target user mentions and as shown in previous researches [83, 91] these Twitterers have an influence on the target user's topical preferences and interests.

- Equation 5.3 provides a model for the probability distribution of words used by those Twitterers whom the target user retweets; as shown by Cha et al [40] that the retweet feature of Twitter serves as an information diffusion mechanism with homophily having a pronounced role and it is for this reason one can hypothesize that Twitterers retweeted by the target user provide a reflection of his/her interests.

- Equation 5.4 provides a model for the probability distribution of words used by those Twitterers whom the target user follows. As De Choudhury [51] shows, there exists a high degree of homophily on Twitter with respect to the follow relationship and it is for this reason one can hypothesize that Twitterers followed by the target user provide a reflection of his/her interests.

Herein, we point out a practical consideration that arises during application of the proposed model. This consideration is with respect to re-computation of the above probabilities within the model (i.e., in Equations 5.1- 5.4) on the arrival of a new tweet. This is computationally feasible only for large-scale search engine companies such as Google, Yahoo! and Bing; we overcome this practical limitation by means of keeping the model static once tweets are downloaded.

5.2.3 Similarity measures between users

The previous subsections presented the individual Twitter models that we propose for the purpose of user profiling. An essential component of the models is the similarity measure *sim(u,u$_i$)* between the target user and the user in his/her Twitter network (more specifically, the user in *mention*, *retweet* or *following* network). As mentioned in section 5.1.1 of this Chapter, a binary inclusion factor (we use *bif* from this point onwards) is an essential component of the model as it is able to make one of the following decisions:

- In case *bif* is on, the similarity measure *sim(u,u$_i$)* is used to either include or exclude a particular Twitter user from within the target user's *mention*, *retweet* or *following* network. This inclusion or exclusion is performed by means of an empirically-set threshold for the value of the similarity measure (i.e., *sim(u,u$_i$)*).
- In case *bif* is off, the similarity measure *sim(u,u$_i$)* is used to weigh a particular Twitter user from within the target user's *mention*, *retweet* or *following* network.

In mathematical terms, *bif* is used to tune the similarity measure *sim(u,u$_i$)* in the following manner:

$$sim(u,u_i) = \begin{cases} 1 & bif = on \text{ and } sim(u,u_i) \geq threshold_{sim} \\ 0 & bif = on \text{ and } sim(u,u_i) \leq threshold_{sim} \\ sim_{nw}(u,u_i) & bif = off \text{ and network similarity to be used} \\ sim_{topical}(u,u_i) & bif = off \text{ and topical similarity to be used} \end{cases}$$

We propose two classes of similarity measures based on the following intuitions:

- Two users are more likely to have common preferences and interests if they share many users within their Twitter networks and hence, we propose network-based similarity measures.
- Two users are more likely to have common preferences and interests if they share interests in the same topics and hence, we propose topical similarity measures.

5.2.3.1 Network-Based Similarity

Previously, we defined U_m as the set of users mentioned by u, U_r as the set of users whose tweets were retweeted by user u and U_f as the set of users whose tweets were followed by user u. We present a network-based similarity measure which we then use as a weighting heuristic for a particular user in U_m, U_r or U_f.

We calculate the similarity between the current user u and each user u_i occurring in either U_m, U_r or U_f based on the heuristic that the more people u_i follows in these sets, the more likely that the current user's interests overlap with the user u. Furthermore, we normalise this score by the maximum of total number of users that user u_i follows or the number of users in U_m, U_r or U_f. We use the following formula to calculate the similarity score between user u and a user $u_i \in U_m$.

$$sim_{nw}(u, u_i) = \frac{|follow(u_i) \cap follow(u)|}{max(|follow(u_i)|, |follow(u)|)} = \frac{|follow(u_i) \cap U_m|}{max(|follow(u_i)|, |U_m|)}$$

where *follow(u_i)* is the set of users followed by u_i.

We also calculate similarity for all users in U_r and U_f using the same approach.

5.2.3.2 Topical Similarity

For the definition of topical similarity we make use of the Twitter-LDA model introduced by Zhao et al. [202] explained in Section 2.3 of Chapter 2 of this thesis. As explained in section 2.3, Twitter-LDA works on the assumption that an author (in this case Twitter user) has an underlying topic distribution. For the purpose of our user model this assumption makes sense in that a particular Twitter user would have topical interests within a fixed set of domains and it is these domains he/she will tweet about. As an example, consider Figure 5.2 again; the biographies of the three Twitter users clearly indicate that there will be a fixed amount of topics related to the interests of these users and Twitter-LDA is able to capture these interests in an effective manner.

We use the Twitter-LDA to determine the tweets' topics from tweets of all the users in sets U_m, U_r and U_f. These topics are are then utilized in a probabilistic model to determine topical similarity between a target user and a user u_i in U_m, U_r or U_f as follows:

$$sim_{topical}(u, u_i) = \frac{\sum_{topic_j \in Topic_{u_i} \cap Topic_u} n(topic_j, Topic_u) + \mu \frac{n(topic_j, Topic_{U_m})}{|Topic_{U_m}|}}{|t_u| + \mu}$$

where *n($topic_j$,$Topic_u$)* denotes the number of tweets by the target user u related to *$topic_j$* and *$topic_j$* denotes a topic that is common between topics from within tweets of target user u and a user u_i in U_m. Moreover, $n(topic_j, Topic_{U_m})$ denotes the number of tweets by users in set U_m related to $topic_j$, and t_u denotes the total number of tweets by the target user u[2]. The topical similarity measure is essentially a weighted average of the commonality between topical distributions of a target user and the users in his/her network and is hence a good indication of shared preferences and interests.

We also calculate similarity for all users in U_r and U_f using the same approach.

5.3 Summary

In this chapter, we presented details of our user profiling strategy which is based on the content generated by a user in a microblog environment. Furthermore, the model also takes into account content generated by those within a user's Twitter network (specifically, his/her followed, mention and retweet network). The model is built upon a statistical language modelling approach which is able to gather long-term user-related information in an implicit manner. This chapter presented an overview of our model while outlining its novel features, namely ability of the model to 1) include features from within a user's Twitter network, 2) incorporate a binary inclusion factor that either includes/excludes similar users or weighs them into the model, and 3) apply network-based and topical similarity measures over the Twitter network of a user. This was followed by explanation of model details whereby individual constituents of the model were explained. Specifically, the model captures the probability distributions of words used by the target user's own tweets, tweets of Twitterers whom the target user mentions, tweets of tweets of Twitterers whom the target user retweets, and tweets of Twitterers whom the target user follows. Finally, we presented details of the two types of similarity measures of which the first takes into account the set of users commonly followed by the target user and the user in his/her Twitter network while the second takes into account common topics shared by the target user and the user in his/her Twitter network.

[2] Note that as before μ is the Dirichlet prior with standard value of '2000'.

Chapter 6

Application of Twitter-based User Model in Web Search Personalization

This chapter describes the application of the proposed Twitter-based user model in a Web search personalization framework. We follow a strategy in which non-personalized search results returned from a search system are re-ranked by means of a user profile defined on the basis of the approach presented in Chapter 5. We begin by describing the overall architecture of the proposed personalized search strategy. This is followed by an explanation of the result adaptation framework which combines the individual Twitter models for re-ranking search results obtained from a non-personalized baseline; we also explain the various settings and heuristics for the model's parameters. Finally, we present details of the various experimental evaluations we undertook for demonstrating effectiveness of the proposed model.

6.1 Overall Architecture of Personalized Search System

Figure 6.1 presents an overall architecture of the personalized search system. It comprises two essential components, namely, the Twitter-based user model presented in Chapter 5 (shown in the bottom half of Figure 6.1 below the block labeled "user profile module") and the retrieval component (shown in the top half of Figure 6.1 above the block labeled "user profile module").

Here, the target user u issues a query to the retrieval module; the retrieved search results are fed to the re-ranking module which uses the computations from the target user's Twitter model, i.e. $P(\ w \mid T\ )$ to re-rank the search results. As explained

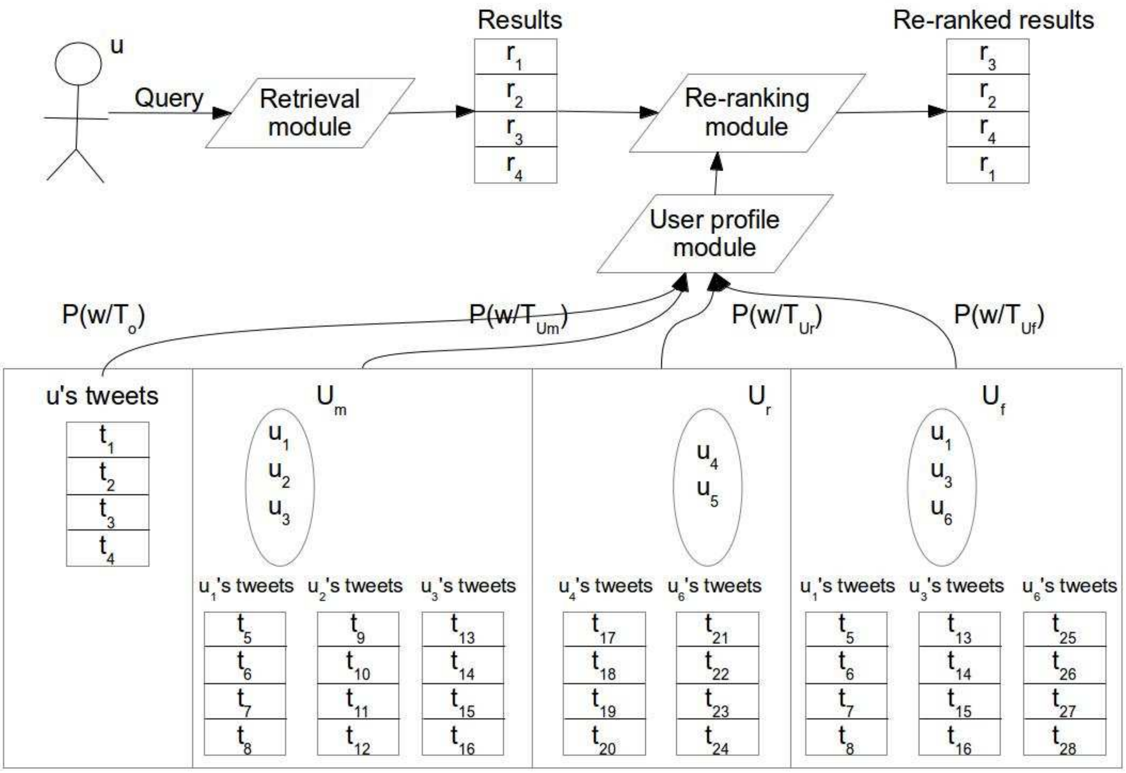


Figure 6.1: Overall Architecture of Personalized Search System

previously the user profile module is built using the user's Twitter content and the content of his/her Twitter network as shown in Figure 6.1. The illustration in the figure demonstrates the target user having four tweets i.e., t_1 - t_4; the distribution $P(w \mid T_o)$ is computed using these tweets and the result of the computation is used by the user profile module. The target user has three users in their mention network U_m i.e., u_1 - u_3 with u_1's tweets being t_5 - t_8, u_2's tweets being t_9 - t_{12} and u_3's tweets being t_{13} - t_{16}; the distribution $P(w \mid T_{U_m})$ is computed using these tweets and the result of the computation is used by the user profile module. The target user has two users in its retweet network U_r i.e., u_4 and u_5 with u_4's tweets being t_{17} - t_{20} and u_5's tweets being t_{21} - t_{24}; the distribution $P(w \mid T_{U_r})$ is computed using these tweets and the result of the computation is used by the user profile module. The target user contains three users in its follow network U_f i.e., u_1, u_3 and u_6 with u_1's tweets being t_5 - t_8, u_3's tweets being t_{13} - t_{16} and u_6's tweets being t_{25} - t_{28}; the distribution $P(w \mid T_{U_f})$ is computed using these tweets and the result of the computation is used by the user profile module. Note that the follow network and the mention network contain the common users u_1 and u_3. The user profile module is the component that computes the probabilistic distributions of the words within the language employed in tweets t_1 - t_{28} and it aids the re-ranking module in computation of personalized scores for each search result. We present further details of how the user profile module works in the next section.

6.2 Overall Model for Web Search Personalization: Re-ranking Framework

The most popular approaches that employ language modelling in information retrieval [136, 158] employ a query-likelihood function which incorporates a ranking function in the retrieval module of the information retrieval system. This ranking function fundamentally estimates the likelihood of the query given a language model for each document in the collection. To the best of our knowledge, with the exception of works in [104, 169] none of the proposed personalized search models in the literature employ language models in the result adaptation step (refer to section 2.4.1.2 of Chapter 2 of this thesis for an overview of "result adaptation"). We aim to achieve this via the Twitter-based user profiles derived through statistical language models as explained in Chapter 5. Fundamentally, our aim is to compute the likelihood of a document's relevance to a user via estimation of the likelihood of the document given a language model for the user.

We now present the overall model for the user that combines the individual Twitter-based user models of the previous chapter. The individual Twitter models based on the target user's own tweets, tweets by Twitterers whom he/she mentions, tweets by Twitterers whom he/she retweets and tweets by Twitterers whom he/she follows are linearly combined to represent an overall model *P(w | T)* as follows:

$$P(w \mid T) = \lambda_o * P(w \mid T_o) + \lambda_m * P(w \mid T_{U_m}) + \lambda_r * P(w \mid T_{U_r}) + \lambda_f * P(w \mid T_{U_f}) \tag{6.1}$$

Here, λ_o, λ_m, λ_r and λ_f represent the parameters for weighting the various components of the model. We present a detailed approach for setting these parameters in Section 6.2.1.

For the re-ranking step, we compute the likelihood of generating a document d from the language model estimated from the target user's Twitter model, i.e. *P(w | T)*, as follows:

$$P(u)_{lm}(d/T) = \sum_{w \in W} P(w \mid T)^{n(w,d)} \tag{6.2}$$

where w is a word in the title and snippet of a document returned by a search system (i.e., d), W the set of all the words in the title and snippet of document d, *n(w,d)* the term frequency of w in d, and u is the target user for whom we want to personalize Web search results.

6.2.1 Using Twitter Behavior Heuristics for Parameter Setting

This section outlines the heuristic used in setting the parameters. We adopt a PageRank like approach to the network of users followed by the target user. Furthermore, depending on the amount of mentions and retweets within the tweets of a user we determine *"trust scores"* which are used as parameters for the model (λ_o, λ_m and λ_r in equation 6.1).

6.2.1.1 Random Surfer Behavior on Twitter Network

As explained in Section 5.2 of Chapter 5, the proposed model takes into account those Twitterers' tweets whom the user mentions, whom the user retweets and whom the user follows. However, the likelihood that the Twitterers followed by a target user reflect his/her preferences is low unless the target user mentions or retweets the Twitterer followed[1]. The model already incorporates the mention and retweet network and hence, the likelihood that the target user is interested in a followed Twitterer mimics the *"random surfer model"* where the random surfer gets bored after several mentions and retweets and switches to a random followed Twitterer. Based on this intuition, we propose the following parameterization for λ_o, λ_m, λ_r and λ_f

$$\alpha = \lambda_o + \lambda_m + \lambda_r \tag{6.3}$$

$$1 - \alpha = \lambda_f \tag{6.4}$$

Here, α represents the damping factor which is basically the probability of the target user's interests being reflected by Twitterers mentioned or retweeted. We set the damping factor to 0.85 which is the standard value used by the PageRank algorithm giving the value for λ_f of 0.15.

[1] It is often the case that random acquaintances are also followed on Twitter.

6.2.1.2 Trust Scores based on Tweeting Activities

Users differ in their behavior on Twitter in that some actively engage in conversations through the mention feature while others diffuse information in the form of retweets [31]. These differences in behavior form the basis for *"trust scores"* within our model. The *"trust scores"* measure the proportion of the target user's own tweets, tweets in which he or she engages in the mention activity and tweets in which he or she engages in the retweet activity. More precisely we set the parameters λ_o, λ_m, λ_r as follows

$$\lambda_o = \frac{|t_o|}{|t_m| + |t_r| + |t_o|} * \alpha \tag{6.5}$$

$$\lambda_m = \frac{|t_m|}{|t_m| + |t_r| + |t_o|} * \alpha \tag{6.6}$$

$$\lambda_r = \frac{|t_r|}{|t_m| + |t_r| + |t_o|} * \alpha \tag{6.7}$$

where, t_o represents original tweets by the target user, t_m represents those tweets by the target user in which he/she engages in the mention activity and t_r represents those tweets by the target user in which he/she engages in the retweet activity.

6.3 Evaluation of the Personalized Search Approach

In this section we describe our experimental evaluations that demonstrate the effectiveness of of our proposed approach. There are three objectives in our experiments.

1. We wish to explore whether personalization through a Twitter-based user profile improves search quality over the underlying non-personalized search engine.

2. We wish to evaluate the effect of the binary inclusion factor introduced in Chapter 5 in an attempt to study the benefits of inclusion/exclusion strategy vs. weighting strategy for users similar to the target user (refer to section 5.2.3 of Chapter 5 to recall aspects of binary inclusion factor).

3. We wish to compare the effectiveness of the proposed network-based similarity measures (refer to section 5.2.3.1 of Chapter 5 to recall network-based similarity measures) vs. the topical similarity measures (refer to section 5.2.3.2 of Chapter 5 to recall topical similarity measures) in an attempt to obtain the optimal setting.

4. We wish to analyze the effect of trust scores on the model parameters in an attempt to study how Twitter behavior-based heuristics affect the experimental outcomes.

The first experimental objective ties in with the first research question raised in Chapter 1 (Section 1.2.1), while the remaining experimental objectives tie in with the third research question raised in Chapter 1 (Section 1.2.1).

6.3.1 Experimental Setup

We conducted both offline and online evaluations through the methods explained in section 2.4.4.1 and 2.4.4.2 of Chapter 2 of this thesis. Through offline evaluations, we determine an optimal setting which is then utilized for the online evaluations.

6.3.1.1 Offline Evaluations

Within the offline evaluations, two sets of experiments were conducted with different sets of users. The reason for recruiting different sets of users for both the studies is due to different values for inter-annotator agreement scores across the different sets of users. We perform this step of measuring inter-annotator agreement via Cohen's kappa to ensure the agreement in relevance judgements between different sets of users in the two studies i.e., the user data used in *"CiteData"* by Harpale et al. [77] (refer to 2.4.4.1 of Chapter 2 of this thesis) and the dataset we created. To calculate inter-annotator agreement across relevance judgements provided by Harpale et al. and our recruited users we asked each user to mark as relevant or irrelevant 50 documents per query; we obtain these 50 documents using a BM25 non-personalized search algorithm (refer to section 2.4.4.1 of Chapter 2 of this thesis for an explanation of BM25 algorithm). Note that each user in our study was asked to mark 50 documents across the queries they selected from the short-listed queries reflecting their interests.

Finally, we calculate the Cohen's kappa across the relevance judgements for the short-listed queries; for the purpose of calculating Cohen's kappa we used the relevance judgements by the graduate students of Harpale et al.'s study and the relevance judgements by the users in our study. We obtain an average Cohen's kappa value of 0.86 across all queries and all users for the first set of users, and average Cohen's kappa value of 0.43 across all queries and all users for the second set of users. Hence, for the first set of experiments we utilized the relevance judgements from *"CiteData"* while for the second set of experiments we utilized the custom relevance judgements provided by the users themselves. This difference in inter-annotator agreement may be

Experiment Type	Users	Chosen Queries	Relevance Judgements	Average Cohen's Kappa
I	14	8	*CiteData*	0.86
II	84	12	*Custom*	0.43

Table 6.1: Summary of Experimental Settings for Offline Evaluations

attributed to the fact that the first set of users comprised early-stage researchers and PhD students (similar to users recruited in Harpale et al.'s study) while the second set of users comprised more advanced researchers. Note that since the dataset comprises academic articles we recruited Twitter users who are academics with specific, personalized information needs for academic articles.

For the first set of experiments, we recruited 14 active Twitter users with permission to use their Twitter data for the purpose of experimental evaluations. As mentioned previously, we obtained the search queries, their corresponding relevance judgements and underlying corpus (i.e., search documents' collection) from *"CiteData"*. We asked each user who participated in our user-study to select a subset of the queries that were similar to a search query that he/she had issued at some point. Each user was asked to select 10 queries from the 41 queries of the dataset; of these we selected the queries that had been selected by at least three users which amounted to a total of eight unique queries.

For the second set of experiments, we recruited 84 active Twitter users and used their Twitter data for the purpose of experimental evaluations. As mentioned previously, we obtained the search queries and underlying corpus (i.e., search documents' collection) from *"CiteData"*. We asked each user who participated in our user-study to select a subset of the queries that were similar to a search query that he/she had issued at some point. Each user was asked to select 12 queries from the 41 queries of the dataset and the re-ranked results were graded as highly relevant (2), relevant (1) and non-relevant (0)[2]. The fact that relevance judgements provided by users in the second set of experiments are graded may also lead to a low inter-annotator agreement with users in *CiteData* dataset. To summarize, we present the various scenarios being tested within the offline evaluations in Table 6.1; Tables 6.2 and 6.3 also show statistics of users in mention, retweet and follow network for both sets of experiments.

[2] Note that since we decided to utilize custom relevance judgements for the second set of experiments, we also asked this second set of users (i.e., the advanced researchers) to provide graded relevance judgements across the retrieved documents.

Average No. of Users in Mention Network, U_m	26
Average No. of Users in Retweet Network, U_r	89
Average No. of Users in Follow Network, U_f	112

Table 6.2: Statistics for Users in First Set of Offline Experiments

Average No. of Users in Mention Network, U_m	76
Average No. of Users in Retweet Network, U_r	32
Average No. of Users in Follow Network, U_f	103

Table 6.3: Statistics for Users in Second Set of Offline Experiments

For both sets of experiments, we re-rank the top-20 search results obtained through a non-personalized BM25 retrieval model.

We also compare the performance of our proposed personalization model against some personalization systems proposed in the literature. As baseline personalization systems we use the approach by Teevan et al. [172] in addition to the approach by Matthijs and Radlinski [105]. However, we replace the search and browsing history data of these approaches by tweets of the target user and his/her network due to the limitation of such history data not being available.

6.3.1.2 Online Evaluation

Having determined the optimal settings through the offline evaluations, a large-scale online interleaved evaluation was conducted and this was to estimate the performance of our system on real users with real information needs so as to ensure that the results of offline evaluation do not overfit to the dataset. A browser plugin was developed and 16 out of the 84 users who participated in the second set of experiments within the offline evaluations agreed to participate in the online evaluation. We performed the interleaved evaluation over a two-week period for the 16 users and we followed an approach similar to Matthijs and Radlinski [105] (refer to section 2.4.4.2 of Chapter 2 of this thesis for an overview of online interleaved evaluation). Search results from Google were re-ranked and the two rankings i.e., the original one from Google and the one produced after re-ranking by our system were interleaved to ensure that a click at random would be equally likely to be on a result from either ranking.

6.3.2 Experimental Results

We first explain the notation we use for presentation of the experimental results as follows:

- *np* is used to denote results from a non-personalized baseline.
- *pbif.uniform* is used to denote results with "binary inclusion factor *bif*" on (refer to section 5.2.3 of Chapter 5 of this thesis).

Chosen Algo	Measures	
	MAP	$P@10$
np	0.389***	0.567**
$p.uniform_{network}$	0.539*	0.673
$p.uniform_{topical}$	0.543**	0.685*
$pbif.uniform_{network}$	0.503*	0.654**
$pbif.uniform_{topical}$	0.514	0.667***

Note *p<.05, **p<.01, ***p<.001

Table 6.4: Comparison of Retrieval Performance for Proposed Personalization Model against Non-Personalized Baseline and across Various Settings for *"bif"*

The subscript *network* is used to denote that the network-based similarity measure of section 5.2.3 is being used and the subscript *topical* is used to denote that the topical similarity of section 5.2.3 is being used. Furthermore, the dot '.' is followed by either *uniform* or *trust* to denote parameter settings for the model (i.e., values of λ_o, λ_m and λ_r in equation 6.1) with *uniform* implying uniform parameter values and *trust* implying parameter values based on trust scores derived using Twitter behaviors (refer to Section 6.2.1.2).

- *p.uniform* is used to denote results with the "binary inclusion factor *bif*" is turned off; the subscript *network* denotes network-based similarity measure being used and *topical* denotes topical similarity being used. As before, dot '.' is followed by either *uniform* or *trust* to denote parameter values either based on uniform scoring or based on trust scores.

6.3.2.1 Experiment Type I

Table 6.4 shows results of the experiment that sets out to address the first and second experimental objective. It shows the results for the first set of experiments which involved 14 Twitter users with relevance judgements from *CiteData*. For this set of experiments, we evaluate the performance of our proposed personalization model with various experimental settings using the evaluation metrics of mean average precision (MAP) and precision at top 10 documents (P@10) which respectively measure the systems overall retrieval accuracy and its performance for those documents that are most viewed. The results clearly show that the use of Twitter data for user profile creation is effective and outperforms a non-personalized by a large degree. Furthermore, the experimental outcomes that correspond to toggling between an on/off setting for the binary inclusion factor *bif* show that weighting the similarity of users into the model is more effective than excluding non-similar users. This is partly due to the nature of Twitter users where even a small proportion of similarity may be due to the fact that a certain user mostly posts tweets with irrelevant content[3]; however, weighting in similarity into the model ensures that no information about the target user is lost in terms of content relating to his/her topical interests within his/her network. Finally, as the results show there is not much difference between user similarities based on network structure and common topics (refer to section 5.2.3 for a recap on these similarity measures); we reason on this aspect later in this section.

Table 6.5 shows results of the experiment that sets out to address the third and fourth experimental objective. Again, the results are corresponding to the first set of experiments involving 14 Twitter users with relevance judgements from *CiteData*. As can be seen when using using uniform parameters for λ_o, λ_m and λ_r in equation 6.1, the MAP and P@10 scores are low as compared to when paramters are assigned on basis of Twitter-based trust scores. This makes intuitive sense in that a given user is not equally represented by those he mentions and retweets; moreover, there is a huge difference among users in terms of their tweeting activities with some tweeting a lot while some rarely tweeting. Moreover, topical similarity performs better in the case of using parameters derived from Twitter behavior-based trust scores. This is due to two reasons: 1) trust scores are able to assign the correct amount of importance to the mention and retweet network of a user, and 2) optimal parameter settings for mention and retweet networks make the effect of topical similarities more prominent by distributing appropriately the effect of each word reflecting the target user's interest.

6.3.2.2 Experiment Type II

Table 6.6 shows the results for the second set of experiments which involved 84 Twitter users with custom relevance judgements. Again, this experiment sets out to address the third and fourth experimental objective. For this set of experiments, we evaluate the performance of our proposed personalization model with various experimental settings using the evaluation metrics of mean average precision (MAP), precision at top 10 documents (P@10) and normalized discounted cumulative gain at top 10 documents (NDCG@10) (NDCG@10 reflects the overall ranking positions of relevant/highly relevant documents). The results in Table 6.6 again demonstrate

[3] Note that irrelevance here implies content not representative of the target user's interests.

Chosen	Measures	
Algo	MAP	P@10
$p.uniform_{network}$	0.539*	0.673
$p.uniform_{topical}$	0.543**	0.685*
$p.trust_{network}$	0.623	0.634
$p.trust_{topical}$	0.711	0.667

Note *p<.05, **p<.01, ***p<.001

Table 6.5: Comparison of Retrieval Performance for Proposed Personalization Model across Uniform and Trust Scores-Based Parameter Assignment and across Network and Topical Similarities

Chosen	Measures		
Algo	MAP	P@10	$NDCG$@10
$p.uniform_{network}$	0.564*	0.551**	0.461
$p.trust_{network}$	0.597	0.582*	0.493***
$p.uniform_{topical}$	0.612*	0.608	0.513**
$p.trust_{topical}$	0.651**	0.643*	0.578***
Teevan	0.541***	0.472	0.420*
Matthijs	0.589	0.564**	0.488*

Note *p<.05, **p<.01, ***p<.001

Table 6.6: Comparison of Retrieval Performance for Proposed Personalization Model across Uniform and Trust Scores-Based Parameter Assignment and across Network and Topical Similarities against other Personalization Models

the strength of topical similarity measures which is more pronounced when parameters are derived from trust scores. We list the reasons behind the success of topical similarity measures as follows:

- The nature of Twitter makes it easy to follow a wide range of people and hence common Twitter users among those being followed may not be necessarily due to common interests. Instead, these common users may include celebrities thereby introducing some amount of noise when network similarity measures are used.

- Topical similarities are able to extract topical interests and preferences more effectively thereby giving higher weight to words that reflect common interests between a target user and his/her Twitter network.

- The difference between network and topical similarities is more pronounced with users in the second set of experiments[4]. This is due to the fact that users in our second set of experiments included mature researchers and followed a diverse set of people among themselves (e.g., researchers within same region followed a range of different people not necessarily belonging to academia). On the other hand all users in the first set of experiments mostly followed mature researchers and hence, network similarities between them had values close to topical similarities.

6.3.2.3 Effect of Individual Twitter Networks

It is important to note the potential effect of individual components of the Twitter network, namely, the mention, retweet, and follow network. For the first set of experiments, the retweet network outperformed the mention and follow network while for the second set of experiments, the mention network outperformed the retweet and follow network[5]. Again, the reason for this difference arises from the nature of the users in the two experiments wherein the first set of experiments was undertaken with early-stage researchers and PhD students while the second set of experiments was undertaken with mature researchers. It is also worth noting that early-stage researchers and PhD students are highly engaged in the retweet activity while mature researchers engage in a lot of conversations with other researchers of their field; this is also obvious from statistics shown in Tables 6.2 and 6.3. It also follows from this observation that the nature of networks is largely driven by the nature of activities users are involved in, and hence, combination of all three networks leads to optimal settings.

6.3.2.4 Online Evaluation

Finally, for the online interleaved evaluation we obtained a total of 518 queries and of these 489 queries received a click on a search result. Of these 489 queries, 302 (61.8%) queries received higher votes across our personalization model while the remaining 187 (38.2%) received higher votes across the original Google rankings. This again demonstrates the potential for search personalization based on Twitter data to improve the search experience. Note that we only personalize using the variant that performs best in the offline evaluation (i.e., the one denoted by $p.trust_{topical}$).

[4] As mentioned previously in Section 6.3.2.1 in Table 6.4 we are revisiting this.

[5] Note that the combination of these networks still shows the best performance.

6.4 Summary

This chapter presented details of the application scenario whereby the Twitter-based user model is applied to Web search personalization. We began by presenting the overall architecture of the personalized search framework; the user model acquired from user's own tweets and tweets of those in his/her network (i.e., those mentioned, retweeted and followed) form the core of the personalized search architecture (see Figure 6.1). The diagrammatic explanation of the personalized search architecture was followed by presentation of the overall model that combines the individual Twitter models of Chapter 5 in a linear fashion through four parameters (i.e., λ_o, λ_m, λ_r, and λ_f) in equation 6.1 and then utilizing this linear combination in a re-ranking approach over words in returned documents (see equation 6.2). We also explained the heuristics used for parameter setting which are influenced by Twitter behavior; more specifically, for the followed users we propose the concept of a "random surfer" wherein a user gets bored after mentions and retweets to switch to those followed and at the same time "trust scores" are used to model degree of interaction with those in mention and retweet network. Detailed experimental evaluations are presented with two kinds of offline evaluations and an online evaluation. The results demonstrate that weighting in users leads to better performance and makes the effects of topical similarities more pronounced; network-based similarity measures do not show a good performance on account of the information nature of Twitter networks. Finally, online evaluations also demonstrated that our technique performs well in a real-world search scenario.

Chapter 7

Application of Twitter-based User Model in Scientific Articles' Recommendation

In this chapter, we present an approach to utilize the Twitter-based user model to suggest scientific articles of interest to novice researchers. The approach in addition to producing effective recommendations for scientific articles alleviates the cold-start problem and is a step towards elimination of the gap between Twitter and science. We begin by presenting the results of a survey conducted in order to analyze how and why early-stage researchers use Twitter. This is followed by a description of the overall architecture of the proposed scientific articles' recommendation framework[1]. We then explain the process of ranking Twitter users followed by a target user in order to obtain a list of top-k researchers that a user follows. We explain the process of ranking research topics of interest to a target user. Both these rankings are performed in conjunction with DBLP paper titles from within our dataset. Finally, we explain the application of the statistical language model of Chapter 5 to tweets of researchers followed by a target user in attempt to score paper titles and serve them as recommendations to the target user. The chapter concludes with some experimental evaluations that demonstrate the effectiveness of our proposed framework.

7.1 Twitter Usage by Researchers

The underlying intuition behind the use of Twitter data for scholarly paper recommendations is common knowledge that scientists while attending conferences, and/or

[1] The tweets by users a Twitterer follows and titles of scientific articles obtained from crawling DBLP comprise two essential sources of data for the recommendation framework.

while conducting experiments tweet about their experiences and these tweets in turn can serve as a rich source for inferring research interests [75, 101]. We set up an online survey to analyze the habits and motivations of early-stage researchers with respect to their use of Twitter. We advertised the survey via University mailing lists and social media services particularly targeting PhD students in early stages of their career and who had an active Twitter account. Our main motivation was to observe the main motivations behind young researchers' use of Twitter and whether or not they consider it as a valuable resource when it comes to staying up-to-date about latest researches in a particular field. This section presents details of the undertaken survey.

We received 280 responses distributed as follows: 65% were PhD students, 10% MSc students, 8% research assistants, 12% postdoctoral researchers, and 5% lecturers. The average number of years using Twitter among our respondents is 2.57 years. One outcome of this survey is that 93% of the respondents use Twitter to stay up-to-date about latest research developments in their respective fields. Other uses of Twitter involved sharing knowledge about their field of expertise and communication about their research projects; however, these goals were shared by the senior researchers among our respondents with the early-stage researchers mainly using it for learning about new researches through the activity of following other researchers. Finally, 87% of our respondents follow approximately 50-100 researchers on Twitter with the average being 67.2 followed researchers per respondent and 69% consider it as highly beneficial in terms of staying up-to-date with latest research in their fields.

Based on the findings of this initial survey of early-stage researchers we observe that most early-stage researchers are turning to Twitter for discovering experts with research interests similar to theirs and hence, their activities on an open medium such as Twitter can be utilized towards the recommendation of scientific articles. The following sections describe the proposed framework in detail.

7.2 Overall Architecture of Scientific Articles' Recommendation

Figure 7.1 presents an overall architecture of the recommendation system for scientific articles. This section describes the proposed recommendation framework in detail. We follow a content-based filtering strategy in which we utilize tweets of Twitterers that a particular user follows (from this point on we refer to these followed Twitterers as followees) along with titles of scientific articles (the module labelled "DBLP paper

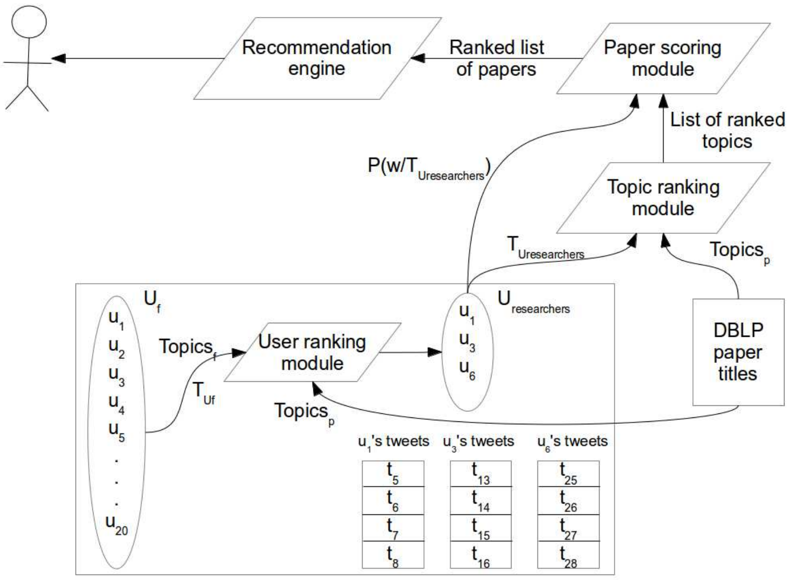


Figure 7.1: Overall Architecture of Recommendation System for Scientific Articles

titles" in Figure 7.1 depicts these titles). Note that we only utilize the model taking into account tweets of followees (we explain the reasoning behind this later in Section 7.3). These users depicted by set U_f are first ranked by the "User ranking module" using topics from within followees' tweets (i.e., $Topics_f$), tweets by followees (i.e., T_{U_f}), and topics from within "DBLP paper titles" (i.e., $Topics_p$) to obtain the set of researchers depicted by set $U_{researchers}$; note that the set $U_{researchers}$ is a subset of U_f.

The illustration in Figure 7.1 shows three researchers in $U_{researchers}$ i.e., u_1, u_3 and u_6; together these tweets comprise the set $T_{U_{researchers}}$. Again, the tweets by researchers $T_{U_{researchers}}$ in conjunction with topics from within "DBLP paper titles" (i.e., $Topics_p$) are used by the "Topic ranking module" to produce a list of ranked topics that represent the target user's topical interests. As can be seen in Figure 7.1, u_1 has tweets t_5 - t_8, u_3 has tweets t_{13} - t_{16}, and u_6 has tweets t_{25} - t_{28}; the distribution $P(w \mid T_{U_{researchers}})$ is computed using these tweets and the result of the computation is used by the "Paper scoring module" to score each paper title representing a topic from within the list of ranked topics. We explain the functioning of "User ranking module" and "Topic ranking module" in Section 7.3, and the functioning of "Paper scoring module" in Section 7.4.

7.3 Filtering Topical Interests corresponding to Paper Titles

This section presents our technique for filtering the research topics that a target user is interested in. As mentioned in Section 7.2, the user profile model that we propose for scientific articles' recommendation is somewhat different from the model proposed in Chapter 5. The difference lies in terms of the model taking into account tweets of those users whom the target users follows (i.e., the followees) while ignoring tweets of users that the target user mentions and retweets (recall these models from Section 5.2.2 of Chapter 5). This is done in context of the particular application scenario of scientific articles' recommendation and on account of the behavior of early-stage researchers on Twitter. There are many early-stage researchers who rarely post content on Twitter i.e., they refrain from the mention and retweet activities on Twitter. However, as the survey results of Section 7.1 show, they follow users representing their research interests with an average of 67.2 researchers followed by a particular user. The following subsections explain the process through which followees are filtered to obtain the top-k researchers (Section 7.3.1), and the process through which topics relevant to a target user's research interests are ranked in order of his/her preference (Section 7.3.2).

7.3.1 Ranking Module for Researchers Followed on Twitter

The first step involves the use of a topic model (more specifically, Twitter-LDA) to obtain the topics from tweets of all the followees of a target user and the paper titles of scientific articles. Twitter-LDA is an unsupervised machine learning technique that discovers latent topics from a Twitter corpus [202]. Twitter-LDA differs from the original LDA framework by Blei et al. [27] in that a single tweet is assigned a single topic instead of a distribution over topics. This is more suited to the task at hand as researchers' when tweeting about their research are very specific and focused, and mostly restricted to one topic. We apply the Twitter-LDA algorithm simultaneously on the followees' tweets and the paper titles with the number of topics set to 200[2]. Our aim is to filter out and produce a ranking of those followees of a user who are involved in scientific research. To this aim, we utilize the intersection of topics found in both paper titles and followees' tweets. Each followee of a user is ranked as follows:

[2] This number is determined empirically after determining the number which clearly distinguishes topics of tweets and paper titles.

$$Rank_{followee} = (\sum_{t \in Topics_p} \frac{n(t, T_{U_f})}{|T_{U_f}|}) * |Topics_p \cap Topics_f| \tag{7.1}$$

where T_{U_f} denotes all tweets by a followee, *Topics*$_p$ denotes the set of topics defining the titles of scientific articles, *Topics*$_f$ denotes the set of topics defining the tweets of a followee and *n(t,*T_{U_f}*)* the number of times a particular topic *'t'* from within *Topics*$_p$ occurs among the tweets of a followee. Accordingly, a followee will be ranked highly if he/she contains a large number of intersections from within topics occurring in titles of scientific articles. Based on the ranking scores of all followees of a particular user, we obtain top-k researchers followed by a target user. The next step involves using the topics from within tweets of these top-k researchers to find the top-n topics of interest to the target user so as to recommend scientific articles from within those topics.

7.3.2 Ranking Module for Research Topics

The scoring framework for topics involves summing up scores for each topic and discovering the dominant topics. Note that we utilize the topics of the set *Topics*$_p$ from within the topics of top-k researchers and this helps avoid noisy topics in the recommendation process. We first determine a score for each topic using the following:

$$Score_{topic} = (\sum_{t \in Topics_p} \frac{n(t, T_{U_{researchers}})}{|T_{U_{researchers}}|}) \tag{7.2}$$

We explain further with the help of the example in Table 7.1. For the sake of understanding the example, assume a total of 10 researchers (i.e., k of top-k researchers equalling 10) and 15 topics t_1-t_{15} with t_3, t_5, t_6, t_7, t_8 and t_{10} belonging to the set *Topics*$_p$. The scores for these topics are combined to produce a final ranking for the topics as shown in Table 7.1; and scientific articles corresponding to these topics are recommended in proportion to the contribution of each topic's score.

For the sake of continuing with this example we take 'n' to be 3 and hence, continue with top-3 topics from *Topics*$_p$ (i.e., t_8, t_{10} and t_7). Note that the incorporation of different topics at this stage enables the generated recommendations to be diverse.

Users	Topics					
	t_3	t_5	t_6	t_7	t_8	t_{10}
u_1	0.13	0.34	0.16	0	0.53	1.13
u_2	0	0.5	0	0.34	0.68	0.43
u_3	0.4	0.12	0.45	0.73	0	0
u_4	0	0.11	0.92	0.22	0	0.64
u_5	0.23	0	0	0.17	0.25	0.55
u_6	0	0.2	0.18	0	1.21	0
u_7	0	0.23	0.38	0.15	0.78	0
u_8	0.48	0	0.14	0.67	0	0.98
u_9	0.19	0	0	0.17	0.93	0
u_{10}	0	0.47	0	0	0.37	0.74

Table 7.1: Example to Illustrate Ranking of Topics Related to Titles of Scientific Articles

	Topics					
	t_3	t_5	t_6	t_7	t_8	t_{10}
Topic Scores	1.43	1.97	2.23	2.45	4.75	4.47

Table 7.2: Final Scores Assigned to Each Topic from within $Topics_p$

7.4 Scoring Framework using Language Models of Followed Researchers

For the purpose of scoring each scientific article, we use a language modelling approach to compute the likelihood of generating an article a from a language model estimated from a user's Twitter followees as follows:

$$P(u)_{t_i}(a/T) = \prod_{w \in a} P(w \mid T)^{n(w,a)} \tag{7.3}$$

where w is a word in the title of articles corresponding to topic t_i (in our example t_i would be t_8, t_{10} and t_7 from among top-3 topics), *n(w,a)* the term frequency of w in a, and u is the user for whom we want to generate the recommendations. Here, T is used to represent the uniform mixture of the Twitter model of researchers followed by a target user as follows:

$$P(w \mid T) = P(w \mid T_{U_{researchers}}) \tag{7.4}$$

Here, $T_{U_{researchers}}$ denotes the tweets by the researchers whom the user u follows corresponding to topic t_i. The Twitter model *P(w* | $T_{U_{researchers}}$*)* can be estimated as:

$$P(w \mid T_{U_{researchers}}) = \frac{1}{|T_{U_{researchers}}|} \sum_{t \in T_{U_{researchers}}} P(w \mid t) \tag{7.5}$$

The constituent language model for $T_{U_{researchers}}$ are a uniform mixture of the language models of researchers' tweets corresponding to topic t_i and employing Dirichlet prior smoothing:

$$P(w \mid t) = \frac{n(w,t) + \mu \frac{n(w, coll)}{|coll|}}{|t| + \mu}$$

where *n(w,.)* denotes the frequency of word w in (.), *coll* is short for collection which refers to all tweets by top-k researchers, and |.| is the overall length of the tweet or the collection. As before, μ is the Dirichlet prior and we utilize the standard value of '2000' for it.

With respect to differences from the model of Chapter 5, another difference arises within the filtering step explained above (refer to Section 7.3); this filtering step replaces the need for any similarity measure of Section 5.2.3 (Chapter 5) being taken into account. Intuitively, application of similarity measures for the task at hand can be prone to errors as similarities may be unable to capture research interests. On the other other hand, performing filtering of researchers and then research topics in conjunction with titles of scientific papers ensures narrowing down of topics relevant to research interests of a target user.

Moreover, equation 7.3 can be modified to include various factors such as freshness score of an academic article (i.e, a measure based upon year of publication), impact factor of venues and/or impact factor of authors. The content-based filtering strategy within our model enables it to be extensible and flexible in addition to being able to produce diverse recommendations. Moreover, the proposed model alleviates the cold-start problem commonly encountered in the recommendation systems' domain whereby user ratings for items to be recommended are not available; in this case however, the tweets by the followees of a target user serve as the starting point.

7.5 Experimental Evaluations

In this section we describe our experimental evaluations that demonstrate the effectiveness of our proposed approach. We first describe the dataset of recruited users along with the dataset of scientific articles followed by details of experimental results.

7.5.1 Experimental Setup

7.5.1.1 Dataset

We recruited 64 active Twitter users with permission to use their Twitter data for the purpose of experimental evaluations. Using the Twitter API, we obtained the tweets of all their followees. Table 7.3 shows some basic statistics about the dataset. The titles of scientific articles are gathered by application of focused crawling to DBLP using the boilerpipe API [95]. A total of 50,252 titles were fetched from a record of various Computer Science conferences and journals from within diverse research fields such as databases, embedded systems, graphics, information retrieval, networks, operating systems, programming languages, software engineering, security, user interface, and social computing.

Average No. of Followees per User	237
Maximum No. of Followees	1022
Minimum No. of Followees	54
Average Tweets per Followee	508
Total Tweets in Collection	32,518

Table 7.3: Statistics about Employed Twitter Dataset

7.5.1.2 Parameters and Evaluation Measures

For the purpose of our experimental evaluations, we set ‘k’ described in Section 7.3 to 30, 60 and 90 respectively i.e., we use top-30, top-60 and top-90 researchers followed by a user for generating his/her list of scientific articles’ recommendation. The number of topics ‘n’ of Section 7.3 is set to 15. As in standard information retrieval, top ranked documents are the most important since users often scan just the first ranks and hence, each user was asked to mark as relevant or irrelevant the top-20 articles recommended to him/her. We evaluated our recommender system using Mean Average Precision (MAP), Mean Reciprocal Rank (MRR) and Precision @ 10 (P@10)[3].

[3] Note that we treat each user as a separate query.

The content-based filtering strategy based on tf-idf by Phelan et al. [132] is used as a baseline to compare the effectiveness of our recommendation model; we explained this strategy in Section 3.4.1 of Chapter 3 of this thesis. We remind the reader that the recommendation system by Phelan et al. suggests a method that promotes news stories from a user's favorite RSS feeds based on Twitter activity through application of a content-based recommendation technique by mining terms from both the RSS feeds and the Twitter messages [132]. Note that to compare our approach with the one proposed by Phelan et al.'s we replace contents in RSS feeds with content of scientific paper titles in order to ensure a fair comparison.

7.5.2 Experimental Results

We evaluate the performance of our proposed recommendation model using the relevance judgements obtained for the 64 users. Table 7.4 shows the experimental results i.e. MAP, MRR and P@10 values for our approach with the different parameter settings for 'k' and for the approach by Phelan et al.; we use student's t-test to verify the soundness of our evaluations and the results corresponding to our model are statistically significant with $p < 0.05$. We report the results together across the judgements for all 64 users.

Our recommendation model built on top of the framework presented in Chapter 5 is able to outperform the tf-idf baseline and this is due to terms introducing a significant amount of noise when recommending scientific articles. On the other hand, topics tend to be clean and give a better representation of research interests of a novice researcher; furthermore, to illustrate the diversity of topics we have shown top-7 words corresponding to top-5 topics for one of the users in our evaluations in Table 7.5. It is evident from Table 7.5 that the topics are from within diverse and non-overlapping research areas and this demonstrates the strength of the topic filtering approach explained in Section 7.3.

On account of applying the language modelling framework of Section 7.4 to tweets derived from within topics, the recommendation process is able to avoid noisy terms and captures research interests accurately. As an example, the approach by Phelan et al. recommends many papers from within the field of "operating systems" to early-stage researchers in "information retrieval" due to the tf-idf score of the term "systems" being high; however, our approach is able to avoid this limitation due to application of topical filtering as explained in Section 7.3[4]. Finally, as evident from

[4] Our framework applies term scoring from within the language modelling framework but restricts it to tweets of researchers belonging to a specific topic t_i.

Chosen Algo	Measures		
	MAP	*MRR*	*P*@10
top-30	0.461	0.667	0.512
top-60	0.651	0.878	0.681
top-90	0.511	0.728	0.643
Phelan et al.	0.384	0.528	0.496

Table 7.4: Comparison of Retrieval Performance for our Proposed Personalization Model

experimental results in Table 7.4 the model with parameter 'k' set to 60 outperforms all the other versions and intuitively this makes sense due to a limited amount of researchers the user is actually interested in (as the survey from Section 7.1 shows that an average of 67.2 researchers are followed by a particular user).

Topic	**Top Terms Describing Topic (from Twitter-LDA)**
Social Network Analysis	social, networks, analysis, online, information, community, detection
Computational Intelligence	fuzzy, neural, networks, control, systems, learning, forecasting
Wireless Networking	network, wireless, sensor, mobile, scheduling, routing, traffic
Information Science	information, query, interfaces, users, relevance, contextual, exploratory
Logic Programming	logic, probabilistic, reasoning, constraint, inference, model, temporal

Table 7.5: Top-7 Words from Top-5 Topics for a User in Experimental Evaluations

7.6 Summary

Researchers are actively turning to Twitter in an attempt to network with other researchers, and stay updated with respect to various scientific breakthroughs. Young and novice researchers have also found Twitter as a valuable source of information in terms of staying up-to-date with various developments in their field of research. In this chapter, we presented an approach built upon our Twitter-based user model of Chapter 5 to utilize this valuable information source for suggesting scientific articles of interest to novice researchers. We applied basic modifications to the model presented in Chapter 5 in an attempt to tailor it to the task at hand thereby demonstrating the strength of the model. The proposed extensions involved taking into account tweets of users followed by a target user while ignoring tweets of users mentioned and retweeted; this is done to reflect the behavior of early-stage researchers on Twitter whereby they do not indulge in mention or retweet activities. Our findings in this chapter presented an attempt to answer the second and third research question in Section 1.2.1 of Chapter 1 of this thesis. We presented a diagrammatic explanation of the overall architecture of our scientific articles' recommendation framework followed by an explanation of the process through which we obtain topics relevant to the target user's research interests. Another modification to the model of Chapter 5 was ignoring similarity measures between a target user and his/her Twitter network and instead utilization of a topic modelling-based filtering strategy to first obtain a list of top-k researchers followed by a list of top-n research topics. The tweets of top-k researchers representing top-n research topics are utilized in the language modelling framework of Chapter 5 to score each paper title (obtained from our custom "DBLP dataset"). The approach in addition to producing effective recommendations for scientific articles alleviated the cold-start problem and served as a step towards elimination of the gap between Twitter and science.

Chapter 8

Conclusion

This chapter provides a broad summary of our work in this thesis. We begin by summarizing the main contributions to the research field in Section 8.1. We then review our findings with respect to our research questions in Section 8.2 followed by summarizing the significance of research outcomes of this thesis in Section 8.3. Finally, in Section 8.4 we discuss future directions for the research conducted in this thesis.

8.1 Summary of Contributions

There is an explosion of information on the World Wide Web today which has lead towards the "information overload" problem . The nature of this problem is getting severe day-by-day; personalization has been extensively explored as a solution to this problem context with most efforts at personalization making use of log-based data for making inferences about users on the World Wide Web. The changing nature of the Web however makes necessary the application of modern means through which users can be modelled and their interests, goals, context, knowledge and preferences captured: this thesis is a significant contribution with regards to utilization of such modern means for modelling users. Specifically, we have contributed to the demonstration of Social Web data as an effective means for making inferences about a user. We presented the notion of "social breadcrumbs" to refer to traces left behind by users during their interactions on modern-day social networking platforms; and attempted to research into the usefulness of these traces by means of creating user profiles with the help of these traces.

Following presents a focussed summary of the contributions of this thesis:

- We have attempted to advance state-of-the-art with respect to the "privacy-personalization" paradox [19, 195]. To the best of our knowledge, this work presents the first attempt at pursuing a study of users' privacy concerns from the viewpoint of their Social Web activities. We conducted extensive user surveys in an attempt to study the correlations between users' privacy concerns with respect to Web search personalization[1] and their social network usage patterns.

  The participants' responses to the survey questions enabled us to use a regression model for identifying the relationship between SNS variables and willingness to personalize Web search. The findings of the model reveal that males are more likely to consider Web search personalization as beneficial while location and age do not have much effect on willingness for Web search personalization. Furthermore, the presence of a user on Twitter and/or Google+ is a strong indicator that he/she will consider Web search personalization as beneficial and similar is the case for his/her high usage of Twitter and/or Google+; the increase is more significant for user presence on Google+ and for high usage of Google+. Additionally, as expected an increase in posting frequency on Facebook increases the odds of willingness to personalize Web search and contrary to expectation, an increase in the number of users' friends on Facebook decreases the odds of willingness to personalize Web search. Lastly, users who use SNS more frequently for Q & A activities are more likely to be willing to opt for Web search personalization in addition to users who frequently consider that responses coming from SNS as more reliable than responses from search engines. With respect to a user's awareness with regards to the personalization feature in current Web search engines, the likelihood is increased if the user is a male, has a presence on LinkedIn, has a high posting frequency on Facebook, and has a high number of tweets on Twitter. With regards to user awareness with regards to search engines making use of user search history data for the process of Web search personalization, the likelihood is increased for males, for users with Twitter and/or LinkedIn presence, and for highly frequent Twitter users. With regards to comfort level with search engines utilization of user search history data for the process of Web search personalization, the likelihood is increased for males, for users of Twitter, Google+ and/or bookmarking sites, for more frequent Google+ users, for users with less mentions and/or high number of tweets on Twitter, for users who frequently use SNS for Q & A activities and

[1] This was taken up as one of the case-studies for personalized applications.

consider responses from SNS to be more reliable than responses coming from search engines.

We also performed a follow-up user survey for use in a support vector machine (SVM) based prediction framework. The prediction results lead to the observation that SNS features such as a user's demographic factors (such as age, gender, location), a user's presence or absence on Twitter and Google+, amount of activity on Twitter and Google+ along with the user's tendency to ask questions on social networks are significant predictors in characterising users who would be willing to opt for personalized Web search results.

- We proposed a statistical language model by means of which user's long-term interests are captured in an effective manner. Two unique aspects of the model are its utilization of an alternate source of data from microblogs (specifically, Twitter) and its ability to take into account various features of a user's Twitter network and his/her behavior on Twitter. More specifically, the model aims to compute the likelihood of an item's relevance to a user via estimation of the likelihood of the item given a language model for the user[2]. Within the user language model, aspects of a user's Twitter posts and his/her Twitter network (users followed, mentioned, retweeted) are used to capture his/her interests and preferences. Moreover, the model also incorporates similarity measures between a user and his/her network based upon user interactions on Twitter i.e., network-based similarity based upon common users followed by a given user and a user in his/her network; or topical similarity based upon the topics present within their tweets.

- The strength of the model is demonstrated by means of its application in two case-studies, namely, Web search personalization, and scientific articles' recommendation. Various extensions and enhancements are applied to the model to adjust it for the application scenarios under consideration; in this context the contributions are as follows:

  - In the application of the proposed Twitter-based user model to personalized search, we introduce the concept of a "random surfer" in which we mimic the behavior of the random surfer where he/she gets bored after several mentions and retweets thereby switching to a random followed

[2] Items in this context can be Web documents retrieved by a Web search engine if the model is applied for Web search personalization, or books to be bought from a book web site if the model is applied for a book recommendation system.

Twitterer in his/her network. Moreover, the concept of "trust scores" is also introduced which captures the level of user engagement via user's own tweeting behavior, his/her mentions and his/her retweets. These Twitter behavior heuristics are utilized for setting the parameters of our Twitter-based user model. Experimental results clearly demonstrate that incorporation of Twitter behavior heuristics into the model shows superior performance.

- In the application of the proposed Twitter-based user model to scientific articles' recommendation, we only take into account tweets of those a user follows while ignoring mention and retweet networks[3]. Furthermore, instead of utilizing similarity measures from within the model we adopt a filtering approach that utilizes topic modelling. The filtering approach is able to obtain researchers followed by Twitter along with topics relevant to his/her research interest; the statistical language model scoring mechanism is then applied tweets representative of the ranked list of research topics to generate recommendations of scientific articles.

8.2 Answers to Research Questions

Let us now examine how our work answers the research questions stated in Chapter 1.

RQ1. ***Can a user's social network usage patterns serve as a window into his/her privacy concerns with respect to personalization?***
This question has been answered by means of the user survey conducted in Chapter 4 whereby we investigated the correlation between users' social network usage patterns and his/her privacy concerns with respect to personalization. Our correlation analysis yielded useful insights in terms of how presence and high usage of various social networking platforms has an effect on personalization willingness and comfort level with search history data. In fact, the most positive correlation was exhibited for users with a presence on Google+ and Twitter along with high usage of these two social networking platforms. Furthermore, Q & A activity on social networking sites is also a strong indicator of openness to personalization. The conclusions derived from the answer to this particular research question forms a basis for our user modelling strategy in this thesis.

[3] Note that the model is flexible in that such a modification is easily incorporated.

RQ2. ***Can sources other than tags and Web page annotations be utilized as a source of evidence for user profile creation?***

This question forms the basis of our user modelling efforts, and was initially derived by the findings of the user survey in Chapter 4. Recall from survey findings in Table 4.2 that only 15.8% of participants in our survey utilize social bookmarks/tags and hence, its feasibility as a source of evidence for user profile creation is questionable. On the other hand, an increasing number of users are turning to microblogging platforms, particularly Twitter. Twitter is unique in that it provides users not only with the opportunity to create their own content but also allows them to engage in rich interactions with other Twitter users. We explored the possibility of utilizing Twitter for user profile creation and proposed a modelling strategy based on statistical language models. The strategy was tuned for application in Web search personalization together with a scientific articles' recommendation framework, and experimental outcomes demonstrate the usefulness of Twitter as a source for effective user modelling.

RQ3. ***Is it possible to utilize the social network information (friendship links) of a user in order to create a richer and more enhanced user profile?***

The answer to this question lies in our exploration of a user's Twitter network within the user model of Chapter 5. The primary modes of user interaction on Twitter are following, mentions and retweets; and our model incorporates these relationships by taking into account the language of users within these networks. The effectiveness of the model is increased by incorporating the content generated by other users to whom the user being modelled is connected in the social network. This is because in many cases a user does not post his/her own tweets but follows, mentions or retweets other Twitter users. We explore incorporation of these users into our model via network-based similarity measures and topical similarity measures; and topical similarity measures show perform better. Furthermore, the degree of interaction with a user is also taken into account via "trust scores" in the parameter setting step during the application of our model to Web search personalization. Our findings reveal that the effect of incorporating the notion of interaction degree and topical similarities with connected users ensures effective performance of personalized applications.

8.3 Significance of Research Outcome

The benefits of approaching the "privacy-personalization" paradox from a social network usage viewpoint are two-fold; first, it helps in characterizing users who are comfortable with personalization and second, it helps in identification leading to use of an effective source that does not entail a huge amount of privacy concerns. In fact, through exploitation of Social Web data, predictions can be made about user preferences without requiring explicit user input thereby solving the problems encountered in explicit user modelling (refer to Section 3.2.1 of Chapter 3 for an overview of these problems). Moreover, utilizing Twitter as a source of evidence for user profile creation lays the necessary groundwork for alleviation of users' privacy concerns with respect to personalization. This is on account of the fact that most Twitter profiles are open in nature. Finally, the work in this thesis serves as a significant basis with regards to efforts aiming towards incorporation of *social information* in the information-seeking process [63].

8.4 Future Directions

There are different research directions generated by the work in this thesis. We list some of these as follows:

- ***Combination of explicit and implicit user modelling approaches:*** An interesting aspect with regards to the user survey on social network usage patterns and personalization-related privacy concerns is its ability to infer users who are comfortable with accumulation of their log data. A future direction could be integration of such a survey within a Web search engine in an attempt to explicitly gather information about users' privacy concerns.

- ***Combination of Social Web data with log-based data:*** Another interesting direction could be merging log-based usage information that is typically acquired by personalized applications with Social Web data in an attempt to create richer user profiles. An implicit outcome of such an effort would be removal of noise from within log-based data.

- ***Utilization of Twitter-based language model in query adaptation framework:*** So far the model was applied in a result adaptation framework with respect to Web search personalization, i.e., we re-ranked search results obtained from a non-personalized baseline. An interesting aspect worth exploring would be application of user model in a query adaptation framework for enhanced Web search personalization.

- ***Integrating concepts within Twitter-based user model:*** The Twitter-based user model operates on set of terms from within documents and tweets; an aspect worth exploring is utilization of concepts from within the tweets for a richer user profile representation.

Appendix A

Survey

Correlating Web Search Personalization to Users Social Network Us

This study is part of a joint research activity at National University of Ireland, Galway and University of Milano-Bicocca. The study focuses on discovering a correlation between Web Search personalization and Social Network usage based on user preferences. Web search personalization can be defined as production of search results centered around user's preferences. For example, when a user issues the query 'apple' on a search engine, he/she might be looking for the company 'Apple' or the fruit 'apple', this ambiguity could be minimized by understanding the user's usual preferences and context.
Note: This survey is purely for academic purposes, the data would contribute as statistics in the study while the users contributing in it shall remain strictly confidential and treated as anonymous (in order to respect user privacy).

Your Name?*

____

Your Twitter handle will also be acceptable. Although we are asking for names we will not publish any names whatsoever; results shall remain anonymous.

Your Gender?*

□ Male □ Female

Your Location?*

____

Your Age?*

□ 1-20 □ 21-30 □ 31-40 □ 41-50 □ Above 50

Your Profession?*

□ Student □ Full-Time Worker □ Family Maker □ Unemployed □ Retired

You can check more than one options if you belong to multiple domains.

Which of the following social networks do you use the most?*

□ Facebook □ Twitter □ Bookmarking sites (Reddit, del.icio.us, Digg etc.)

□ Google+ □ LinkedIn □ Other: ______________________________

Which of the following social network accounts do you have?*

□ Facebook □ Twitter □ Bookmarking sites (Reddit, del.icio.us, Digg etc.)

□ Google+ □ LinkedIn □ Other: ______________________________

What is your frequency of usage of the most used social network that you chose above?*

□ Several times a day □ Once or twice a day □ Weekly □ Monthly

□ Every few months □ Yearly □ Never

How often do you post something on Facebook (sharing link, photo, status update etc. ?*

□ Frequently (Almost daily or sometimes more than once per day)

□ Sometimes (After several days) □ Rarely □ Never

Choose never if you do not have a Facebook account.

How often do you like something on Facebook (it can be any post or page)?*

□ Frequently (Almost daily or sometimes more than once per day)

□ Sometimes (After several days) □ Rarely □ Never

Choose never if you do not have a Facebook account.

Give the approximate number of your Facebook friends:*

□ Less than 100 □ 100-200 □ 201-300 □ 301-400 □ 401-500

□ More than 500

Select only if you have a Facebook account.

Web Search Personalization Preferences

This section of the survey concerns Web search personalization feature offered by search engines of today and user's level of interaction with this feature.

Are you aware of the personalization feature in Web search engines such as iGoogle?*

□ Yes □ No □ Don't Know

Do you think that personalized search results would be a benefit to you?*

□ Yes □ No □ Don't Know

Are you aware that search engines make use of your search history data (clickthrough data or query logs) for personalization?*

□ Yes □ No

Clickthrough data refers to history of clicked links across queries and query logs refers to a complete log of queries you enter on search engine.

Are you comfortable with the fact that search engines analyze your search activity (history) data for the sake of personalizing Web search results?*

□ Yes □ No

Do you agree with the notion that personalization in Web search engines makes the process of information-seeking easier and less painstaking?*

□ Strongly Agree □ Agree □ Neither Agree nor Disagree □ Disagree
□ Strongly Disagree

QA Behavior on Social Networks

This section of the survey concerns question and answering patterns on social networks.

Have you ever used social networks for information-seeking?*
□ Yes □ No
Information-seeking is a general name given to the activity of posing queries to your friends/acquaintances on social networks for example seeking their opinions on consumer products or services, or seeking advice.

Question-answering activity on social networks is a useful task for you?*
□ Yes □ No

How often do you ask questions on social networks?*
□ Most of the time □ Sometimes □ Rarely □ Never

How often do you feel that responses coming from social networks are more reliable than those coming from Web search engines?*
□ Most of the time □ Sometimes □ Rarely □ Never
Please respond to this question if you have ever used social networks for information seeking.

Appendix B

Publications

Following are the list of papers which form the main body of the thesis.

- Younus, A., Qureshi, M. A., Manchanda, P., ORiordan, C., & Pasi, G. (2014). Utilizing microblog data in a topic modelling framework for scientific articles recommendation. In Social Informatics (pp. 384-395). Springer International Publishing.

- Younus, A., ORiordan, C., & Pasi, G. (2014). A Language Modeling Approach to Personalized Search Based on Users Microblog Behavior. In Advances in Information Retrieval (pp. 727-732). Springer International Publishing.

- Younus, A., ORiordan, C., & Pasi, G. (2013). Predictors of users willingness to personalize web search. In Flexible Query Answering Systems (pp. 459-470). Springer Berlin Heidelberg.

Other papers published during thesis — related topics in *Information Retrieval*, *Social Media Analytics* or *Online Education* but not central to this thesis.

- Younus, A., Qureshi, M.A., Griffith, J., O'Riordan, C., & Pasi, G. (2015). A Study into the Correlation between Narcissism and Facebook Communication Patterns. In Web Intelligence Conference.

- Qureshi, M. A., Younus, A., Yousuf, M., Moiz, A., Saeed, M., Touheed, N., ... & Pasi, G. (2014). YummyKarachi: Using Real-Time Tweets for Restaurant Recommendations in an Unsafe Location. In UMAP Workshops.

- Younus, A., Qureshi, M. A., Saeed, M., Touheed, N., O'Riordan, C., & Pasi, G. (2014, April). Election trolling: analyzing sentiment in tweets during pakistan

elections 2013. In Proceedings of the companion publication of the 23rd international conference on World wide web companion (pp. 411-412). International World Wide Web Conferences Steering Committee.

- Younos, A. (2012). Online education for developing contexts. ACM Crossroads, 19(2), 27-29.
- Younus, A., O'Riordan, C., & Pasi, G. (2012). CIRGDISCO at RepLab2012 Filtering Task: A Two-Pass Approach for Company Name Disambiguation in Tweets. In CLEF (Online Working Notes/Labs/Workshop).
- Younus, A., Qureshi, M. A., Kingrani, S. K., Saeed, M., Touheed, N., O'Riordan, C., & Gabriella, P. (2012, April). Investigating bias in traditional media through social media. In Proceedings of the 21st international conference companion on World Wide Web (pp. 643-644). ACM.

Bibliography

[1] Roy Morgan Research: Privacy and the Community. Prepared for the Office of the Federal Privacy Commissioner, Sydney. `http://aiweb.techfak.uni-bielefeld.de/content/bworld-robot-control-software/`, 2001.

[2] Fifth Study of the Internet by the Digital Future Project Finds Major New Trends in Online Use for Political Campaigns, Center for the Digital Future, Annenberg School, University of Southern California, 2005. [Online; accessed 29-June-2015].

[3] Yankee: Interactive Consumers in the Twenty-First Century: Emerging Online Consumer Profiles, Access Strategies and Application Usage, Yankee Group 23 Oct. 2001.

[4] Fabian Abel, Ilknur Celik, Geert-Jan Houben, and Patrick Siehndel. Leveraging the semantics of tweets for adaptive faceted search on twitter. In *The Semantic Web–ISWC 2011*, pages 1–17. Springer, 2011.

[5] Fabian Abel, Qi Gao, Geert-Jan Houben, and Ke Tao. Analyzing temporal dynamics in twitter profiles for personalized recommendations in the social web. In *ACM WebSci'11*, pages 1–8, June 2011. WebSci Conference 2011.

[6] Fabian Abel, Qi Gao, Geert-Jan Houben, and Ke Tao. Analyzing user modeling on twitter for personalized news recommendations. In *User Modeling, Adaption and Personalization*, pages 1–12. Springer, 2011.

[7] Fabian Abel, Qi Gao, Geert-Jan Houben, and Ke Tao. Semantic enrichment of twitter posts for user profile construction on the social web. In Grigoris Antoniou, Marko Grobelnik, Elena Simperl, Bijan Parsia, Dimitris Plexousakis, Pieter De Leenheer, and Jeff Pan, editors, *The Semanic Web: Research and Applications*, volume 6644 of *Lecture Notes in Computer Science*, pages 375–389. Springer Berlin Heidelberg, 2011.

[8] Fabian Abel, Eelco Herder, Geert-Jan Houben, Nicola Henze, and Daniel Krause. Cross-system user modeling and personalization on the social web. *User Modeling and User-Adapted Interaction*, 23(2-3):169–209, 2013.

[9] Mark S Ackerman, Lorrie Faith Cranor, and Joseph Reagle. Privacy in e-commerce: examining user scenarios and privacy preferences. In *Proceedings of the 1st ACM conference on Electronic commerce*, pages 1–8. ACM, 1999.

[10] Lada A Adamic. The information life of social networks. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, pages 273–274. ACM, 2015.

[11] Deepak Agarwal and Bee-Chung Chen. flda: Matrix factorization through latent dirichlet allocation. In *Proceedings of the Third ACM International Conference on Web Search and Data Mining*, WSDM '10, pages 91–100, New York, NY, USA, 2010. ACM.

[12] Eugene Agichtein, Eric Brill, and Susan Dumais. 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*, SIGIR '06. ACM.

[13] Rakib Ahmed and Shuk Ying Ho. Privacy concerns of users for location-based mobile personalization. In *CONF-IRM 2011 Proceedings*, 2011.

[14] Jae-wook Ahn, Peter Brusilovsky, Jonathan Grady, Daqing He, and Sue Yeon Syn. Open user profiles for adaptive news systems: help or harm? In *Proceedings of the 16th international conference on World Wide Web*, pages 11–20. ACM, 2007.

[15] Sarabjot Singh Anand and Bamshad Mobasher. Contextual recommendation, from web to social web: Discovering and deploying user and content profiles: Workshop on web mining, webmine 2006, Berlin, Germany, september 18, 2006. revised selected and invited papers, 2007.

[16] Annie Anton, Julia B Earp, Jessica D Young, et al. How internet users' privacy concerns have evolved since 2002. *Security & Privacy, IEEE*, 8(1):21–27, 2010.

[17] Robert M. Arlein, Ben Jai, Markus Jakobsson, Fabian Monrose, and Michael K. Reiter. Privacy-preserving global customization. In *Proceedings of the 2Nd ACM Conference on Electronic Commerce*, EC '00, pages 176–184, 2000.

[18] Fabio A Asnicar and Carlo Tasso. ifweb: a prototype of user model-based intelligent agent for document filtering and navigation in the world wide web. In *Sixth International Conference on User Modeling*, pages 2–5, 1997.

[19] Naveen Farag Awad and MS Krishnan. The personalization privacy paradox: an empirical evaluation of information transparency and the willingness to be profiled online for personalization. *MIS quarterly*, pages 13–28, 2006.

[20] Ricardo Baeza-Yates, Berthier Ribeiro-Neto, et al. *Modern information retrieval*, volume 463. ACM press New York, 1999.

[21] Shenghua Bao, Guirong Xue, Xiaoyuan Wu, Yong Yu, Ben Fei, and Zhong Su. Optimizing web search using social annotations. In *Proceedings of the 16th international conference on World Wide Web*, WWW '07, pages 501–510, 2007.

[22] Lamjed Ben Jabeur. *Leveraging social relevance: using social networks to enhance literature access and microblog search.* PhD thesis, Université de Toulouse, Université Toulouse III-Paul Sabatier, 2013.

[23] Paul N Bennett, Filip Radlinski, Ryen W White, and Emine Yilmaz. Inferring and using location metadata to personalize web search. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, pages 135–144. ACM, 2011.

[24] Paul N Bennett, Ryen W White, Wei Chu, Susan T Dumais, Peter Bailey, Fedor Borisyuk, and Xiaoyuan Cui. Modeling the impact of short-and long-term behavior on search personalization. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, pages 185–194. ACM, 2012.

[25] S. Bergamaschi, F. Guerra, and B. Leiba. Guest editors' introduction: Information overload. *Internet Computing, IEEE*, 14(6):10–13, Nov 2010.

[26] Shlomo Berkovsky, Yaniv Eytani, Tsvi Kuflik, and Francesco Ricci. Privacy-enhanced collaborative filtering. In *Proc. User Modeling Workshop on Privacy-Enhanced Personalization*, 2005.

[27] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003.

[28] Jesús Bobadilla, Fernando Ortega, Antonio Hernando, and Abraham Gutiérrez. Recommender systems survey. *Knowledge-Based Systems*, 46:109–132, 2013.

[29] Mohamed Reda Bouadjenek, Hakim Hacid, and Mokrane Bouzeghoub. Laicos: An open source platform for personalized social web search. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '13, pages 1446–1449, 2013.

[30] Mohamed Reda Bouadjenek, Hakim Hacid, Mokrane Bouzeghoub, and Athena Vakali. Using social annotations to enhance document representation for personalized search. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '13, pages 1049–1052, 2013.

[31] Danah Boyd, Scott Golder, and Gilad Lotan. Tweet, tweet, retweet: Conversational aspects of retweeting on twitter. In *System Sciences (HICSS), 2010 43rd Hawaii International Conference on*, pages 1–10. IEEE, 2010.

[32] Danah M Boyd and Eszter Hargittai. Facebook privacy settings: Who cares? *First Monday*, 15(8), 2010.

[33] John S Breese, David Heckerman, and Carl Kadie. Empirical analysis of predictive algorithms for collaborative filtering. In *Proceedings of the Fourteenth conference on Uncertainty in artificial intelligence*, pages 43–52. Morgan Kaufmann Publishers Inc., 1998.

[34] Peter Brusilovsky, Alfred Kobsa, and Wolfgang Nejdl. *The adaptive web: methods and strategies of web personalization*, volume 4321. Springer Science & Business Media, 2007.

[35] Peter Brusilovsky and Carlo Tasso. Preface to special issue on user modeling for web information retrieval. *User Modeling and User-Adapted Interaction*, 14(2):147–157, 2004.

[36] Robin Burke. Hybrid recommender systems: Survey and experiments. *User modeling and user-adapted interaction*, 12(4):331–370, 2002.

[37] Robin Burke. Hybrid web recommender systems. In *The adaptive web*, pages 377–408. Springer, 2007.

[38] John Canny. Collaborative filtering with privacy. In *Security and Privacy, 2002. Proceedings. 2002 IEEE Symposium on*, pages 45–57. IEEE, 2002.

[39] David Carmel, Naama Zwerdling, Ido Guy, Shila Ofek-Koifman, Nadav Har'el, Inbal Ronen, Erel Uziel, Sivan Yogev, and Sergey Chernov. Personalized social search based on the user's social network. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, CIKM '09, pages 1227–1236, 2009.

[40] Meeyoung Cha, Hamed Haddadi, Fabricio Benevenuto, and P Krishna Gummadi. Measuring user influence in twitter: The million follower fallacy. *ICWSM*, 10(10-17):30, 2010.

[41] Ramnath K Chellappa and Raymond G Sin. Personalization versus privacy: An empirical examination of the online consumers dilemma. *Information Technology and Management*, 6(2-3):181–202, 2005.

[42] Jilin Chen, Rowan Nairn, Les Nelson, Michael Bernstein, and Ed Chi. Short and tweet: experiments on recommending content from information streams. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1185–1194. ACM, 2010.

[43] Liren Chen and Katia Sycara. Webmate: a personal agent for browsing and searching. In *Proceedings of the second international conference on Autonomous agents*, pages 132–139. ACM, 1998.

[44] Peng Chen, Huafeng Xie, Sergei Maslov, and Sidney Redner. Finding scientific gems with googles pagerank algorithm. *Journal of Informetrics*, 1(1):8–15, 2007.

[45] Paul Alexandru Chirita, Claudiu S. Firan, and Wolfgang Nejdl. Personalized query expansion for the web. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '07, pages 7–14, 2007.

[46] Paul Alexandru Chirita, Wolfgang Nejdl, Raluca Paiu, and Christian Kohlschütter. Using odp metadata to personalize search. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 178–185. ACM, 2005.

[47] Kevyn Collins-Thompson, Paul N Bennett, Ryen W White, Sebastian de la Chica, and David Sontag. Personalizing web search results by reading level. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, pages 403–412. ACM, 2011.

[48] Vlad Coroama and Marc Langheinrich. Personalized vehicle insurance rates-a case for client-side personalization in ubiquitous computing. In *in Ubiquitous Computing. Workshop on Privacy-Enhanced Personalization. CHI 2006*. Citeseer, 2006.

[49] Mary J Culnan. The culnan-milne survey on consumers & online privacy notices: Summary of responses. In *Interagency Public Workshop: Get Noticed: Effective Financial Privacy Notices*, pages 47–54, 2001.

[50] Mariam Daoud, Lynda Tamine-Lechani, and Mohand Boughanem. Learning user interests for a session-based personalized search. In *Proceedings of the second international symposium on Information interaction in context*, pages 57–64. ACM, 2008.

[51] Munmun De Choudhury. Tie formation on twitter: Homophily and structure of egocentric networks. In *Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on*, pages 465–470. IEEE, 2011.

[52] Gianmarco De Francisci Morales, Aristides Gionis, and Claudio Lucchese. From chatter to headlines: harnessing the real-time web for personalized news recommendation. In *Proceedings of the fifth ACM international conference on Web search and data mining*, pages 153–162. ACM, 2012.

[53] Ying Ding and Jing Jiang. Extracting interest tags from twitter user biographies. In *Information Retrieval Technology*, pages 268–279. Springer, 2014.

[54] David DiSalvo. Are social networks messing with your head? *Scientific American Mind*, 20:48–55, January 2010.

[55] Anlei Dong, Ruiqiang Zhang, Pranam Kolari, Jing Bai, Fernando Diaz, Yi Chang, Zhaohui Zheng, and Hongyuan Zha. Time is of the essence: improving recency ranking using twitter data. In *Proceedings of the 19th international conference on World wide web*, WWW '10, pages 331–340, 2010.

[56] Anlei Dong, Ruiqiang Zhang, Pranam Kolari, Jing Bai, Fernando Diaz, Yi Chang, Zhaohui Zheng, and Hongyuan Zha. Time is of the essence: improving recency ranking using twitter data. In *Proceedings of the 19th international conference on World wide web*, pages 331–340. ACM, 2010.

[57] Zhicheng Dou, Ruihua Song, and Ji-Rong Wen. A large-scale evaluation and analysis of personalized search strategies. WWW '07, pages 581–590, 2007.

[58] Yajuan Duan, Long Jiang, Tao Qin, Ming Zhou, and Heung-Yeung Shum. An empirical study on learning to rank of tweets. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 295–303. Association for Computational Linguistics, 2010.

[59] Khalid El-Arini and Carlos Guestrin. Beyond keyword search: Discovering relevant scientific literature. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '11, pages 439–447, New York, NY, USA, 2011. ACM.

[60] Ali Mamdouh Elkahky, Yang Song, and Xiaodong He. A multi-view deep learning approach for cross domain user modeling in recommendation systems. In *Proceedings of the 24th International Conference on World Wide Web*, pages 278–288. International World Wide Web Conferences Steering Committee, 2015.

[61] Gunther Eysenbach. Can tweets predict citations? metrics of social impact based on twitter and correlation with traditional metrics of scientific impact. *Journal of medical Internet research*, 13(4), 2011.

[62] Wei Feng and Jianyong Wang. Retweet or not?: personalized tweet re-ranking. In *Proceedings of the sixth ACM international conference on Web search and data mining*, pages 577–586. ACM, 2013.

[63] Jennifer Fernquist and Ed H Chi. Perception and understanding of social annotations in web search. In *Proceedings of the 22nd international conference on World Wide Web*, pages 403–412. International World Wide Web Conferences Steering Committee, 2013.

[64] Susannah Fox, Lee Rainie, John Horrigan, Amanda Lenhart, Tom Spooner, and Cornelia Carter. Trust and privacy online: Why americans want to rewrite the rules. *The Pew Internet & American Life Project*, 2000.

[65] Dan Frankowski, Dan Cosley, Shilad Sen, Loren Terveen, and John Riedl. You are what you say: privacy risks of public mentions. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 565–572. ACM, 2006.

[66] Wei Gao, Cheng Niu, Jian-Yun Nie, Ming Zhou, Jian Hu, Kam-Fai Wong, and Hsiao-Wuen Hon. Cross-lingual query suggestion using query logs of different languages. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 463–470. ACM, 2007.

[67] Susan Gauch, Mirco Speretta, Aravind Chandramouli, and Alessandro Micarelli. User profiles for personalized information access. In *The adaptive web*, pages 54–89. Springer, 2007.

[68] Simon Gerber, Michael Fry, Judy Kay, Bob Kummerfeld, Glen Pink, and Rainer Wasinger. *PersonisJ: mobile, client-side user modelling.* Springer, 2010.

[69] M Rami Ghorab, Dong Zhou, Alexander OConnor, and Vincent Wade. Personalised information retrieval: survey and classification. *User Modeling and User-Adapted Interaction*, 23(4):381–443, 2013.

[70] Fréderic Godin, Viktor Slavkovikj, Wesley De Neve, Benjamin Schrauwen, and Rik Van de Walle. Using topic models for twitter hashtag recommendation. In *Proceedings of the 22nd international conference on World Wide Web companion*, pages 593–596. International World Wide Web Conferences Steering Committee, 2013.

[71] Miha Grčar, Blaž Fortuna, Dunja Mladenič, and Marko Grobelnik. knn versus svm in the collaborative filtering framework. In *Data Science and Classification*, pages 251–260. Springer, 2006.

[72] Ramanathan Guha, Vineet Gupta, Vivek Raghunathan, and Ramakrishnan Srikant. User modeling for a personal assistant. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, pages 275–284. ACM, 2015.

[73] Pankaj Gupta, Ashish Goel, Jimmy Lin, Aneesh Sharma, Dong Wang, and Reza Zadeh. Wtf: The who to follow service at twitter. In *Proceedings of the 22nd international conference on World Wide Web*, pages 505–514. International World Wide Web Conferences Steering Committee, 2013.

[74] Ido Guy, Naama Zwerdling, Inbal Ronen, David Carmel, and Erel Uziel. Social media recommendation based on people and tags. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '10, pages 194–201, 2010.

[75] Asmelash Teka Hadgu and Robert Jäschke. Identifying and analyzing researchers on twitter. In *Proceedings of the 2014 ACM Conference on Web Science*, WebSci '14, pages 23–32, New York, NY, USA, 2014. ACM.

[76] John Hannon, Kevin McCarthy, and Barry Smyth. Finding useful users on twitter: twittomender the followee recommender. In *Advances in Information Retrieval*, pages 784–787. Springer, 2011.

[77] Abhay Harpale, Yiming Yang, Siddharth Gopal, Daqing He, and Zhen Yue. Citedata: a new multi-faceted dataset for evaluating personalized search performance. CIKM '10, pages 549–558, New York, NY, USA, 2010. ACM.

[78] Jonathan L Herlocker, Joseph A Konstan, Loren G Terveen, and John T Riedl. Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems (TOIS)*, 22(1):5–53, 2004.

[79] Paul Heymann, Georgia Koutrika, and Hector Garcia-Molina. Can social bookmarking improve web search? WSDM '08, pages 195–206, 2008.

[80] Michael Hitchens, Judy Kay, Bob Kummerfeld, and Ajay Brar. Secure identity management for pseudo-anonymous service access. In *Security in Pervasive Computing*, pages 48–55. Springer, 2005.

[81] Donna L Hoffman, Thomas P Novak, and Marcos Peralta. Building consumer trust online. *Communications of the ACM*, 42(4):80–85, 1999.

[82] Thomas Hofmann. 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*, pages 259–266. ACM, 2003.

[83] Courtenay Honey and Susan C Herring. Beyond microblogging: Conversation and collaboration via twitter. In *System Sciences, 2009. HICSS'09. 42nd Hawaii International Conference on*, pages 1–10. IEEE, 2009.

[84] Damon Horowitz and Sepandar D. Kamvar. The anatomy of a large-scale social search engine. In *Proceedings of the 19th international conference on World wide web*, WWW '10, pages 431–440, New York, NY, USA, 2010. ACM.

[85] Shanshan Huang and Xiaojun Wan. Akminer: Domain-specific knowledge graph mining from academic literatures. In Xuemin Lin, Yannis Manolopoulos, Divesh Srivastava, and Guangyan Huang, editors, *Web Information Systems Engineering WISE 2013*, volume 8181 of *Lecture Notes in Computer Science*, pages 241–255. Springer Berlin Heidelberg, 2013.

[86] Harris Interactive. A survey of consumer privacy attitudes and behaviors. *Rochester, NY*, 47, 2000.

[87] Ameni Kacem, Mohand Boughanem, and Rim Faiz. Time-sensitive user profile for optimizing search personlization. In Vania Dimitrova, Tsvi Kuflik, David Chin, Francesco Ricci, Peter Dolog, and Geert-Jan Houben, editors, *User Modeling, Adaptation, and Personalization*, volume 8538 of *Lecture Notes in Computer Science*, pages 111–121. 2014.

[88] Paul B Kantor, Lior Rokach, Francesco Ricci, and Bracha Shapira. *Recommender systems handbook*. Springer, 2011.

[89] Judy Kay. Scrutable adaptation: because we can and must. In *Adaptive hypermedia and adaptive web-based systems*, pages 11–19. Springer, 2006.

[90] Judy Kay and Gord McCalla. Coming of age: Celebrating a quarter century of user modeling and personalization: Guest editors introduction. *User Modeling and User-Adapted Interaction*, 22(1):1–7, 2012.

[91] Suin Kim, JinYeong Bak, and Alice Haeyun Oh. Do you feel what I feel? social aspects of emotions in twitter conversations. In *ICWSM*, 2012.

[92] Alfred Kobsa. User modeling in dialog systems: Potentials and hazards. *AI & society*, 4(3):214–231, 1990.

[93] Alfred Kobsa. Privacy-enhanced web personalization. In Peter Brusilovsky, Alfred Kobsa, and Wolfgang Nejdl, editors, *The Adaptive Web*, volume 4321 of *Lecture Notes in Computer Science*, pages 628–670. Springer Berlin Heidelberg, 2007.

[94] Alfred Kobsa and Jörg Schreck. Privacy through pseudonymity in user-adaptive systems. *ACM Transactions on Internet Technology (TOIT)*, 3(2):149–183, 2003.

[95] Christian Kohlschütter, Peter Fankhauser, and Wolfgang Nejdl. Boilerplate detection using shallow text features. In *Proceedings of the third ACM international conference on Web search and data mining*, pages 441–450. ACM, 2010.

[96] Yehuda Koren. Factorization meets the neighborhood: a multifaceted collaborative filtering model. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 426–434. ACM, 2008.

[97] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. What is twitter, a social network or a news media? In *Proceedings of the 19th international conference on World wide web*, pages 591–600. ACM, 2010.

[98] Su Mon Kywe, Tuan-Anh Hoang, Ee-Peng Lim, and Feida Zhu. On recommending hashtags in twitter networks. In *Social Informatics*, pages 337–350. Springer, 2012.

[99] Jung-Hyun Lee, Jongwoo Ha, Jin-Yong Jung, and Sangkeun Lee. Semantic contextual advertising based on the open directory project. *ACM Transactions on the Web (TWEB)*, 7(4):24, 2013.

[100] Julie Letierce, Alexandre Passant, John Breslin, and Stefan Decker. Understanding how Twitter is used to widely spread Scientific Messages. In *Proceedings of the WebSci10: Extending the Frontiers of Society On-Line*, March 2010.

[101] Julie Letierce, Alexandre Passant, John G Breslin, and Stefan Decker. Using twitter during an academic conference: The# iswc2009 use-case. In *ICWSM*, 2010.

[102] Kenneth Wai-Ting Leung, Dik Lun Lee, and Wang-Chien Lee. Personalized web search with location preferences. In *Data Engineering (ICDE), 2010 IEEE 26th International Conference on*, pages 701–712. IEEE, 2010.

[103] Fang Liu, Clement Yu, and Weiyi Meng. Personalized web search for improving retrieval effectiveness. *Knowledge and Data Engineering, IEEE transactions on*, 16(1):28–40, 2004.

[104] Julia Luxenburger, Shady Elbassuoni, and Gerhard Weikum. Matching task profiles and user needs in personalized web search. In *Proceedings of the 17th ACM conference on Information and knowledge management*, pages 689–698. ACM, 2008.

[105] Nicolaas Matthijs and Filip Radlinski. Personalizing web search using long term browsing history. WSDM '11, pages 25–34, 2011.

[106] Andrew McCallum, Xuerui Wang, and Andrés Corrada-Emmanuel. Topic and role discovery in social networks with experiments on enron and academic email. *Journal of Artificial Intelligence Research*, pages 249–272, 2007.

[107] Aleecia McDonald and Lorrie Faith Cranor. Beliefs and behaviors: Internet users' understanding of behavioral advertising. TPRC, 2010.

[108] Sean M. McNee, Istvan Albert, Dan Cosley, Prateep Gopalkrishnan, Shyong K. Lam, Al Mamunur Rashid, Joseph A. Konstan, and John Riedl. On the recommending of citations for research papers. In *Proceedings of the 2002 ACM Conference on Computer Supported Cooperative Work*, CSCW '02, pages 116–125, New York, NY, USA, 2002. ACM.

[109] Rishabh Mehrotra, Scott Sanner, Wray Buntine, and Lexing Xie. Improving lda topic models for microblogs via tweet pooling and automatic labeling. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, pages 889–892. ACM, 2013.

[110] Qiaozhu Mei and Kenneth Church. Entropy of search logs: how hard is search? with personalization? with backoff? In *Proceedings of the 2008 International Conference on Web Search and Data Mining*, pages 45–54. ACM, 2008.

[111] Miriam J Metzger. Privacy, trust, and disclosure: Exploring barriers to electronic commerce. *Journal of Computer-Mediated Communication*, 9(4), 2004.

[112] Alessandro Micarelli, Fabio Gasparetti, Filippo Sciarrone, and Susan Gauch. Personalized search on the world wide web. In *The adaptive web*, pages 195–230. Springer, 2007.

[113] Alessandro Micarelli and Filippo Sciarrone. Anatomy and empirical evaluation of an adaptive web-based information filtering system. *User Modeling and User-Adapted Interaction*, 14(2-3):159–200, 2004.

[114] Bradley N Miller, Joseph A Konstan, and John Riedl. Pocketlens: Toward a personal recommender system. *ACM Transactions on Information Systems (TOIS)*, 22(3):437–476, 2004.

[115] J Mills et al. Factors determining the performance of indexing systems. *Volume I-Design, Volume II-Test Results, ASLIB Cranfield Project, Reprinted in Sparck Jones & Willett, Readings in Information Retrieval*, 1966.

[116] Alan L Montgomery and Michael D Smith. Prospects for personalization on the internet. *Journal of Interactive Marketing*, 23(2):130–137, 2009.

[117] Meredith R. Morris, Teevan Jaime, and Katrina Panovich. A Comparison of Information Seeking Using Search Engines and Social Networks. ICWSM '10, pages 291–294, 2010.

[118] Meredith Ringel Morris and Jaime Teevan. Exploring the complementary roles of social networks and search engines. In *Human-Computer Interaction Consortium Workshop (HCIC)*, 2012.

[119] Meredith Ringel Morris, Jaime Teevan, and Katrina Panovich. What do people ask their social networks, and why?: a survey study of status message q & a behavior. CHI '10, pages 1739–1748, New York, NY, USA, 2010. ACM.

[120] Maurice D Mulvenna, Sarabjot S Anand, and Alex G Büchner. Personalization on the net using web mining: introduction. *Communications of the ACM*, 43(8):122–125, 2000.

[121] Cataldo Musto, Giovanni Semeraro, Pasquale Lops, and Marco de Gemmis. Combining distributional semantics and entity linking for context-aware content-based recommendation. In *User Modeling, Adaptation, and Personalization*, pages 381–392. Springer, 2014.

[122] Abhinay Nagpal, Sudheendra Hangal, Rifat Reza Joyee, and Monica S Lam. Friends, romans, countrymen: lend me your urls. using social chatter to personalize web search. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*, pages 461–470. ACM, 2012.

[123] Ellen Nakashima. Aol search queries open window onto users worlds. *The Washington Post*, 2006.

[124] Quang Nhat Nguyen and Francesco Ricci. Long-term and session-specific user preferences in a mobile recommender system. In *Proceedings of the 13th International Conference on Intelligent User Interfaces*, IUI '08, pages 381–384, 2008.

[125] Michael G. Noll and Christoph Meinel. Web search personalization via social bookmarking and tagging. ISWC'07/ASWC'07, pages 367–380, 2007.

[126] Fabrizio Orlandi. *Profiling user interests on the social semantic web.* PhD thesis, 2014.

[127] Fabrizio Orlandi, John Breslin, and Alexandre Passant. Aggregated, interoperable and multi-domain user profiles for the social web. In *Proceedings of the 8th International Conference on Semantic Systems*, pages 41–48. ACM, 2012.

[128] Georgios Paliouras. Discovery of web user communities and their role in personalization. *User Modeling and User-Adapted Interaction*, 22(1-2):151–175, 2012.

[129] Saurabh Panjwani, Nisheeth Shrivastava, Saurabh Shukla, and Sharad Jaiswal. Understanding the privacy-personalization dilemma for web search: a user perspective. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 3427–3430. ACM, 2013.

[130] Michael J Pazzani and Daniel Billsus. Content-based recommendation systems. In *The adaptive web*, pages 325–341. Springer, 2007.

[131] Maria Soledad Pera and Yiu-Kai Ng. With a little help from my friends: Generating personalized book recommendations using data extracted from a social website. In *Proceedings of the 2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology-Volume 01*, pages 96–99. IEEE Computer Society, 2011.

[132] Owen Phelan, Kevin McCarthy, and Barry Smyth. Using twitter to recommend real-time topical news. In *Proceedings of the third ACM conference on Recommender systems*, pages 385–388. ACM, 2009.

[133] Joseph Phelps, Glen Nowak, and Elizabeth Ferrell. Privacy concerns and consumer willingness to provide personal information. *Journal of Public Policy & Marketing*, 19(1):27–41, 2000.

[134] D. Pierrakos and G. Paliouras. Personalizing web directories with the aid of web usage data. *Knowledge and Data Engineering, IEEE Transactions on*, 22(9):1331–1344, Sept 2010.

[135] James Pitkow, Hinrich Schütze, Todd Cass, Rob Cooley, Don Turnbull, Andy Edmonds, Eytan Adar, and Thomas Breuel. Personalized search. *Commun. ACM*, 45(9):50–55, September 2002.

[136] Jay M Ponte and W Bruce Croft. A language modeling approach to information retrieval. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 275–281. ACM, 1998.

[137] Jason Priem and Bradely H Hemminger. Scientometrics 2.0: New metrics of scholarly impact on the social web. *First Monday*, 15(7), 2010.

[138] Feng Qiu and Junghoo Cho. Automatic identification of user interest for personalized search. In *Proceedings of the 15th international conference on World Wide Web*, pages 727–736. ACM, 2006.

[139] Daniele Quercia, Harry Askham, and Jon Crowcroft. Tweetlda: supervised topic classification and link prediction in twitter. In *Proceedings of the 4th Annual ACM Web Science Conference*, pages 247–250. ACM, 2012.

[140] Filip Radlinski, Madhu Kurup, and Thorsten Joachims. How does clickthrough data reflect retrieval quality? In *Proceedings of the 17th ACM conference on Information and knowledge management*, pages 43–52. ACM, 2008.

[141] Daniel Ramage, Susan T Dumais, and Daniel J Liebling. Characterizing microblogs with topic models. *ICWSM*, 10:1–1, 2010.

[142] Daniel Ramage, David Hall, Ramesh Nallapati, and Christopher D Manning. Labeled lda: A supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1*, pages 248–256. Association for Computational Linguistics, 2009.

[143] Josyula R Rao and Pankaj Rohatgi. Can pseudonymity really guarantee privacy? In *USENIX Security Symposium*, 2000.

[144] Francesco Ricci, Dario Cavada, Nader Mirzadeh, and Adriano Venturini. Case-based travel recommendations. *Destination Recommendation Systems: Behavioural Foundations and Applications*, pages 67–93, 2006.

[145] S. E. Robertson and S. Walker. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '94, pages 232–241, 1994.

[146] Stephen E Robertson, Steve Walker, Susan Jones, Micheline M Hancock-Beaulieu, Mike Gatford, et al. Okapi at trec-3. *NIST SPECIAL PUBLICATION SP*, pages 109–109, 1995.

[147] Michal Rosen-Zvi, Thomas Griffiths, Mark Steyvers, and Padhraic Smyth. The author-topic model for authors and documents. In *Proceedings of the 20th conference on Uncertainty in artificial intelligence*, pages 487–494. AUAI Press, 2004.

[148] Roni Rosenfield. Two decades of statistical language modeling: Where do we go from here? 2000.

[149] Matthew Rowe. The credibility of digital identity information on the social web: a user study. In *Proceedings of the 4th workshop on Information credibility*, WICOW '10, pages 35–42, 2010.

[150] Gerard Salton. Automatic information organization and retrieval. 1968.

[151] Amit Sharma and Dan Cosley. Network-centric recommendation: Personalization with and in social networks. In *Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on*, pages 282–289. IEEE, 2011.

[152] Naveen Kumar Sharma. *Discovering topical experts in Twitter social network*. PhD thesis, Indian Institute of Technology, Kharagpur, 2012.

[153] Kim Bartel Sheehan and Mariea Grubbs Hoy. Flaming, complaining, abstaining: How online users respond to privacy concerns. *Journal of advertising*, 28(3):37–51, 1999.

[154] Xuehua Shen, Bin Tan, and ChengXiang Zhai. Privacy protection in personalized search. *SIGIR Forum*, 41(1):4–17, June 2007.

[155] Andriy Shepitsen, Jonathan Gemmell, Bamshad Mobasher, and Robin Burke. Personalized recommendation in social tagging systems using hierarchical clustering. In *Proceedings of the 2008 ACM conference on Recommender systems*, pages 259–266. ACM, 2008.

[156] Milad Shokouhi. Learning to personalize query auto-completion. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, pages 103–112. ACM, 2013.

[157] Barry Smyth and Evelyn Balfe. Anonymous personalization in collaborative web search. *Information retrieval*, 9(2):165–190, 2006.

[158] Fei Song and W Bruce Croft. A general language model for information retrieval. In *Proceedings of the eighth international conference on Information and knowledge management*, pages 316–321. ACM, 1999.

[159] David Sontag, Kevyn Collins-Thompson, Paul N Bennett, Ryen W White, Susan Dumais, and Bodo Billerbeck. Probabilistic models for personalizing web search. In *Proceedings of the fifth ACM international conference on Web search and data mining*, pages 433–442. ACM, 2012.

[160] Micro Speretta and Susan Gauch. Personalized search based on user search histories. In *Web Intelligence, 2005. Proceedings. The 2005 IEEE/WIC/ACM International Conference on*, pages 622–628. IEEE, 2005.

[161] Smitha Sriram, Xuehua Shen, and Chengxiang Zhai. A session-based search engine. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 492–493. ACM, 2004.

[162] Ashok N Srivastava and Mehran Sahami. *Text mining: Classification, clustering, and applications.* CRC Press, 2009.

[163] Sofia Stamou and Alexandros Ntoulas. Search personalization through query and page topical analysis. *User Modeling and User-Adapted Interaction*, 19(1-2):5–33, 2009.

[164] Frederic D Stutzman. *Networked Information Behavior in Life Transition.* PhD thesis, THE UNIVERSITY OF NORTH CAROLINA AT CHAPEL HILL, 2011.

[165] Kazunari Sugiyama, Kenji Hatano, and Masatoshi Yoshikawa. Adaptive web search based on user profile constructed without any effort from users. In *Proceedings of the 13th international conference on World Wide Web*, pages 675–684. ACM, 2004.

[166] Jian-Tao Sun, Hua-Jun Zeng, Huan Liu, Yuchang Lu, and Zheng Chen. Cubesvd: a novel approach to personalized web search. In *Proceedings of the 14th international conference on World Wide Web*, pages 382–390. ACM, 2005.

[167] Latanya Sweeney. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(05):557–570, 2002.

[168] Shayan A Tabrizi, Azadeh Shakery, Masoud Asadpour, Maziar Abbasi, and Mohammad Ali Tavallaie. Personalized pagerank clustering: A graph clustering algorithm based on random walks. *Physica A: Statistical Mechanics and its Applications*, 392(22):5772–5785, 2013.

[169] Bin Tan, Xuehua Shen, and ChengXiang Zhai. Mining long-term search history to improve search accuracy. KDD '06, pages 718–723, 2006.

[170] Lei Tang and Huan Liu. Community detection and mining in social media. *Synthesis Lectures on Data Mining and Knowledge Discovery*, 2(1):1–137, 2010.

[171] Ya Tang. *The design and study of pedagogical paper recommendation.* PhD thesis, University of Saskatchewan, 2008.

[172] Jaime Teevan, Susan T. Dumais, and Eric Horvitz. Personalizing search via automated analysis of interests and activities. SIGIR '05, pages 449–456, 2005.

[173] Jaime Teevan, Susan T. Dumais, and Daniel J. Liebling. To personalize or not to personalize: modeling queries with variation in user intent. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '08, pages 163–170, 2008.

[174] Maximilian Teltzrow and Alfred Kobsa. Impacts of user privacy preferences on personalized systems. In *Designing personalized user experiences in eCommerce*, pages 315–332. Springer, 2004.

[175] Eran Toch, Yang Wang, and Lorrie Faith Cranor. Personalization and privacy: a survey of privacy risks and remedies in personalization-based systems. *User Modeling and User-Adapted Interaction*, 22(1-2):203–220, 2012.

[176] Zeynep Tufekci. Facebook, youth and privacy in networked publics. In *Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media*, volume 148, pages 36–37, 2012.

[177] Joseph Turow, Jennifer King, Chris Jay Hoofnagle, Amy Bleakley, and Michael Hennessy. Americans reject tailored advertising and three activities that enable it, september 2009. In *SSRN: http://ssrn. com/abstract*, volume 1478214.

[178] Yury Ustinovskiy and Pavel Serdyukov. Personalization of web-search using short-term browsing context. In *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management*, pages 1979–1988. ACM, 2013.

[179] David Vallet, Iván Cantador, and Joemon M. Jose. Personalizing web search with folksonomy-based user and document profiles. ECIR'2010, pages 420–431, 2010.

[180] David Vallet, Pablo Castells, Miriam Fernández, Phivos Mylonas, and Yannis Avrithis. Personalized content retrieval in context using ontological knowledge. *Circuits and Systems for Video Technology, IEEE Transactions on*, 17(3):336–346, 2007.

[181] Evelien Van De Garde-Perik, Panos Markopoulos, Boris De Ruyter, Berry Eggen, and Wijnand Ijsselsteijn. Investigating privacy attitudes and behavior in relation to personalization. *Social Science Computer Review*, 26(1):20–43, 2008.

[182] Eduardo Vicente-López, Luis M de Campos, Juan M Fernández-Luna, Juan F Huete, Antonio Tagua-Jiménez, and Carmen Tur-Vigil. An automatic methodology to evaluate personalized information retrieval systems. *User Modeling and User-Adapted Interaction*, 25(1):1–37, 2014.

[183] Graham Vickery and Sacha Wunsch-Vincent. *Participative web and user-created content: Web 2.0 wikis and social networking*. Organization for Economic Co-operation and Development (OECD), 2007.

[184] Chong Wang and David M. Blei. Collaborative topic modeling for recommending scientific articles. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '11, pages 448–456, New York, NY, USA, 2011. ACM.

[185] Qihua Wang and Hongxia Jin. Exploring online social activities for adaptive search personalization. CIKM '10, pages 999–1008, New York, NY, USA, 2010. ACM.

[186] Yang Wang and Alfred Kobsa. Respecting users individual privacy constraints in web personalization. In *User Modeling 2007*, pages 157–166. Springer, 2007.

[187] Ingmar Weber and Carlos Castillo. The demographics of web search. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pages 523–530. ACM, 2010.

[188] Katrin Weller, Evelyn Dröge, and Cornelius Puschmann. Citation analysis in twitter: Approaches for defining and measuring information flows within tweets during scientific conferences. In *# MSM*, pages 1–12, 2011.

[189] Ryen W. White, Peter Bailey, and Liwei Chen. Predicting user interests from contextual information. In *Proceedings of the 32Nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '09, pages 363–370, New York, NY, USA, 2009. ACM.

[190] Ryen W White, Paul N Bennett, and Susan T Dumais. Predicting short-term interests using activity-based search context. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pages 1009–1018. ACM, 2010.

[191] Ryen W White, Wei Chu, Ahmed Hassan, Xiaodong He, Yang Song, and Hongning Wang. Enhancing personalized search by mining and modeling task behavior. In *Proceedings of the 22nd international conference on World Wide Web*, pages 1411–1420. International World Wide Web Conferences Steering Committee, 2013.

[192] Craig E Wills and Can Tatar. Understanding what they do with what they know. In *Proceedings of the 2012 ACM Workshop on Privacy in the Electronic Society*, pages 13–18. ACM, 2012.

[193] Allison Woodruff, Rich Gossweiler, James Pitkow, Ed H. Chi, and Stuart K. Card. Enhancing a digital book with a reading recommender. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '00, pages 153–160, New York, NY, USA, 2000. ACM.

[194] Liang Xiang, Quan Yuan, Shiwan Zhao, Li Chen, Xiatian Zhang, Qing Yang, and Jimeng Sun. Temporal recommendation on graphs via long- and short-term preference fusion. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '10, pages 723–732, New York, NY, USA, 2010. ACM.

[195] Heng Xu, Xin Robert Luo, John M Carroll, and Mary Beth Rosson. The personalization privacy paradox: An exploratory study of decision making process for location-aware marketing. *Decision Support Systems*, 51(1):42–52, 2011.

[196] Shengliang Xu, Shenghua Bao, Ben Fei, Zhong Su, and Yong Yu. Exploring folksonomy for personalized search. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 155–162. ACM, 2008.

[197] Songhua Xu, Hao Jiang, and Francis Chi-Moon Lau. Mining user dwell time for personalized web search re-ranking. In *International Joint Conference on Artificial Intelligence (IJCAI 2011)*, pages 2367–2372. AAAI Press/International Joint Conferences on Artificial Intelligence., 2011.

[198] Chenxing Yang, Baogang Wei, Jiangqin Wu, Yin Zhang, and Liang Zhang. Cares: A ranking-oriented cadal recommender system. In *Proceedings of the 9th ACM/IEEE-CS Joint Conference on Digital Libraries*, JCDL '09, pages 203–212, New York, NY, USA, 2009. ACM.

[199] Arjumand Younus, Colm ORiordan, and Gabriella Pasi. Predictors of users willingness to personalize web search. In *FQAS'13*, volume 8132, pages 459–470. 2013.

[200] Arjumand Younus, Colm ORiordan, and Gabriella Pasi. A language modeling approach to personalized search based on users microblog behavior. In *Advances in Information Retrieval*, pages 727–732. Springer, 2014.

[201] Arjumand Younus, Muhammad Atif Qureshi, Pikakshi Manchanda, Colm ORiordan, and Gabriella Pasi. Utilizing microblog data in a topic modelling framework for scientific articles recommendation. In *Social Informatics*, pages 384–395. Springer, 2014.

[202] Wayne Xin Zhao, Jing Jiang, Jianshu Weng, Jing He, Ee-Peng Lim, Hongfei Yan, and Xiaoming Li. Comparing twitter and traditional media using topic models. ECIR'11, pages 338–349, Berlin, Heidelberg, 2011. Springer-Verlag.

[203] Rong Zheng, Dennis Wilkinson, and Foster Provost. Social network collaborative filtering. *Stern, IOMS Department, CeDER, Vol*, 2008.

[204] Dong Zhou, Séamus Lawless, and Vincent Wade. Improving search via personalized query expansion using social media. *Information retrieval*, 15(3-4):218–242, 2012.

[205] Dong Zhou, Séamus Lawless, and Vincent Wade. Web search personalization using social data. In *Theory and Practice of Digital Libraries*, pages 298–310. Springer, 2012.

[206] Xiangmin Zhou and Lei Chen. Event detection over twitter social media streams. *The VLDB JournalThe International Journal on Very Large Data Bases*, 23(3):381–400, 2014.

[207] C Lawrence Zitnick and Takeo Kanade. Maximum entropy for collaborative filtering. In *Proceedings of the 20th conference on Uncertainty in artificial intelligence*, pages 636–643. AUAI Press, 2004.