User Modeling and Personalization in the Microblogging Sphere

Qi Gao

User Modeling and Personalization in the Microblogging Sphere

Proefschrift

ter verkrijging van de graad van doctor
aan de Technische Universiteit Delft,
op gezag van de Rector Magnificus prof.ir. K.C.A.M. Luyben,
voorzitter van het College voor Promoties,
in het openbaar te verdedigen op maandag 28 oktober 2013 om 15:00 uur

door **Qi GAO**

Bachelor of Engineering in Automation, Tongji University,
geboren te Jiashan, Zhejiang, China.

Dit proefschrift is goedgekeurd door de promotoren:
Prof.dr.ir. G.J.P.M. Houben

Samenstelling promotiecommissie:

Rector Magnificus	voorzitter
Prof.dr.ir. G.J.P.M. Houben	Technische Universiteit Delft, promotor
Prof.dr. P. Brusilovsky	University of Pittsburgh
Prof.dr. P.M.E. De Bra	Technische Universiteit Eindhoven
Prof.dr. V.G. Dimitrova	University of Leeds
Prof.dr. A. Hanjalic	Technische Universiteit Delft
Dr. F. Abel	XING AG
Prof.dr.ir. D.H.J. Epema	Technische Universiteit Delft (reservelid)

SIKS Dissertation Series No. 2013-33

The research reported in this thesis has been carried out under the auspices of SIKS, the Dutch Research School for Information and Knowledge Systems.

Published and distributed by: Qi Gao
E-mail: qigaosh@gmail.com

ISBN: 978-94-6186-227-3

Keywords: user modeling, personalization, recommender systems, semantic web, social web, microblog, twitter, sina weibo

Cover image: Amayzun, “Shell Macro” via Flickr, Creative Commons Attribution.

Printed and bound in The Netherlands by CPI Wörmann Print Service.

Acknowledgments

First and foremost I would like to thank my promotor Prof. Geert-Jan Houben who gave me the opportunity to carry out this PhD in the Web Information Systems (WIS) group. I appreciate the freedom that he has given me in trying new ideas and making my own choices along the way. I want to express my sincere gratitude for his extensive guidance and continuous support to my PhD work.

I am extremely indebted to my advisor Dr. Fabian Abel without whom this thesis would not have been possible. It has been a great pleasure to work with him. Fabian, thank you for the support, the fund and inspiring discussions, and keeping advising me even after you moved to Hamburg.

Besides my promoter and advisor, I would like to thank the rest of my thesis committee: Prof. Peter Brusilovsky, Prof. Paul De Bra, Prof. Vania Dimitrova, Prof. Dick Epema, and Prof. Alan Hanjalic, for their time spent on this thesis and their insightful feedback.

I am grateful for working with the WIS group and other colleagues. I appreciate the help and friendship of Stefano Bocconi, Alessandro Bozzon, Ilknur Çelik, Claudia Hauff, Laura Hollink, Damir Juric, Erwin Leonardi, and Richard Stronkman. I would like to thank my officemate, Jan Hidders, for translating the propositions along with this thesis. I also enjoyed interacting and collaborating with other PhD students: Samur Araújo, Engin Bozdag, Beibei Hu, Jasper Oosterman, Yue Shi, Ke Tao, and Jie Yang. My many thanks go to Rina Abbriata, Ilse Oonk, Franca Post, and Esther van Rooijen for their help and assistant with administrative issues in the past four years. I also want to thank Paulo Anita, Munire van der Kruyk, and Stephen van der Laan for their excellent ICT support.

My special thanks go to my former advisors Prof. Junwei Yan and Dr. Min Liu in Tongji University. I also take this opportunity to thank Prof. Yong Yu, Dr. Haofen Wang, and many other friends in Shanghai Jiaotong University. Their support made my research visit to Shanghai productive and joyful.

I have spent a great time in Delft with many good friends. Thank you all and

take care!

Last but certainly not least, I would like to thank my parents who always encourage me to explore and find my own way. My gratitude to them is beyond words. My utmost gratitude goes to my wife Qin Zhou for her unconditional support and love throughout these years.

Qi Gao
October 2013
Delft

Contents

Acknowledgments **v**

1 Introduction **1**

1.1 Thesis Outline 4

1.2 Origin of Chapters 5

2 Background **7**

2.1 User Modeling 7

2.1.1 Overview 7

2.1.2 User Profiling in the Social Web 9

2.1.3 User Modeling for the Social Semantic Web 14

2.2 Recommender Systems 17

2.2.1 Overview 17

2.2.2 Collaborative Filtering Recommender Systems 18

2.2.3 Content-based Recommender Systems 20

2.3 Research Challenges tackled in this Thesis 22

3 Microblogging-based User Modeling Framework **25**

3.1 Introduction 25

3.2 TweetUM - Tweet-based User Modeling Framework 27

3.2.1 Topic Modeling 29

3.2.2 Enrichment 32

3.2.3 Temporal Constraints 34

3.2.4 Weighting Schemes 35

3.3 GeniUS - Generic User Modeling Library for the Social Semantic Web . . . 37
3.3.1 Architecture of GeniUS . . . 37
3.3.2 Domain-specific User Profile Construction Using GeniUS . 40
3.4 Discussion . . . 43

4 Semantic Enrichment for Microblogging-based User Modeling 47
4.1 Introduction . . . 47
4.2 Exploitation of Linkage for Microblogging-based User Modeling . . 49
4.2.1 Linkage Discovery Strategies . . . 52
4.2.2 Evaluation of Linkage Discovery . . . 55
4.2.3 Analyzing User Profile Construction based on Linkage Discovery . . . 59
4.3 Exploitation of Emotion for Microblogging-based User Modeling . 62
4.3.1 Emotions in Microposts . . . 63
4.3.2 Emotion Classification Strategies . . . 64
4.3.3 Evaluation of Emotion Classification . . . 66
4.3.4 Analyzing Emotion-based User Profiles . . . 68
4.4 Discussion . . . 71

5 Microblogging-based User Modeling for Culture-aware Analytics 73
5.1 Introduction . . . 73
5.2 Analysis of Users' Microblogging Behavior on Sina Weibo and Twitter . . . 75
5.2.1 Methodology . . . 76
5.2.2 Analysis of Access Behavior . . . 79
5.2.3 Syntactic Content Analysis . . . 81
5.2.4 Semantic Content Analysis . . . 84
5.2.5 Sentiment Analysis . . . 86
5.2.6 Analysis of Temporal Behavior . . . 88
5.2.7 Interpretation of Findings . . . 91
5.3 Analysis of Information Propagation on Sina Weibo and Twitter . . 92
5.3.1 Research Questions . . . 92
5.3.2 Reposting Frequency . . . 93
5.3.3 Reposting Speed . . . 94

5.3.4 Broadness of User Interests 94
5.3.5 Syntactical Characteristics of propagated messages 95
5.3.6 Sentiment Characteristics of propagated messages 96
5.3.7 Interpretation of Findings 97
5.4 Discussion 98

6 Microblogging-based User Modeling for Personalized Recommendations 101
6.1 Introduction 101
6.2 Analyzing User Modeling on Twitter for Personalized News Recommendation 104
6.2.1 Analysis of Twitter-based User Profiles 104
6.2.2 Exploitation of User Profiles for Personalized News Recommendations 109
6.2.3 Synopsis 112
6.3 Interweaving Trend and User Modeling on Twitter for Personalized News Recommendation 113
6.3.1 Trend Modeling on Twitter 113
6.3.2 Temporal Analysis of User and Trend Profiles on Twitter . . 117
6.3.3 Evaluation of Trend and User Modeling for Recommending News Articles 120
6.3.4 Synopsis 123
6.4 Analyzing Temporal Dynamic on Twitter for Personalization 124
6.4.1 Evolution of User Interests in Trending Topics 125
6.4.2 Time-sensitive User Modeling for Personalized Recommendations 131
6.4.3 Synopsis 135
6.5 Domain-specific User Modeling on Twitter for Personalized Recommendations 136
6.5.1 Analysis of Domain-Specific User Profile Construction . . . 136
6.5.2 Evaluation of Domain-Specific User Profile Construction for Recommendation System 139
6.5.3 Synopsis 142
6.6 Discussion 143

7 Conclusion 147

7.1 Summary of Contributions 147
7.2 Future Work 152

Bibliography **155**

List of Figures **175**

List of Tables **177**

Summary **179**

Samenvatting **181**

Curriculum Vitae **183**

Chapter 1

Introduction

Throughout the last years, microblogging has become a popular mechanism for information sharing and communication on the Web. For example, Twitter, as the most prominent microblogging service, serves more than 500 million users who post over 340 million short messages every day[1], sharing their thoughts and everyday activities with the public. On microblogging platforms, users are able to post messages, which are limited to a certain maximum length (e.g., 140 characters on Twitter), as well as repost messages of other users. In addition, users can follow other users so that they can receive the latest posts published by those users. Microblogging services such as Twitter also provide APIs that allow third parties to access microblogging data and develop various external applications such as systems for event detection [168, 181], opinion mining [40] or personalized recommendations [44, 85].

As microblogging services have gained immense popularity around the world, more and more people post real-time messages via different devices to discuss a variety of topics. Given the plethora of digital traces that people leave on the microblogging platforms, researchers have started exploiting microblogging activities for understanding users' information needs and modeling users' preferences [28, 96]. Some research initiatives focus on inferring specific attributes of a user from microblogging data such as the user's location [153], political orientation [78], or influential power [42]. However, there are interesting research questions regarding user modeling based on microblogging activities that have not been studied yet. How can we learn the semantics of microblogging activities and infer users' interests from those activities? How can we construct user profiles based on microblogging data to support different applications such as personalized recommender sys-

[1] http://techcrunch.com/2012/07/30/analyst-twitter-passed-500m-users-in-june-2012-140m-of-them-in-us-jakarta-biggest-tweeting-city/

tems? In this thesis, we explore these questions and introduce a generic framework for user modeling based on user and usage data collected from microblogging platforms. Our user modeling framework aims at understanding the semantics of individual microblogging activities and allows for generating semantically meaningful user interest profiles to support different external applications. We analyze several design dimensions in the context of user modeling and develop a variety of solutions that allows for the adaptation of the user modeling process to given applications and circumstances.

Given the variety and recency of topics that people discuss on microblogging platforms, user profiles that are generated from microposts promise to be beneficial for other applications on the Web for objectives such as event detection [168], crisis management [6] or expert mining [75]. Due to the shortness of messages posted via microblogging services, making sense of microblogging activities is however a non-trivial task. There is an urgent need to investigate methods for enriching the semantics of microblogging data so that user profiles constructed with rich semantics can be reused and shared across applications [30, 146]. While research efforts have been invested in exploiting textual features of microposts (e.g., hashtags) to understand trending topics [109] or inferring user interests from microblogging streams [132], analyzing the feasibility of correlating microposts with external Web resources for enriching the semantics of user activities on microblogging platforms has not been researched yet. In this thesis, we introduce and evaluate strategies that exploit external Web resources that are related to microposts. Moreover, we analyze and incorporate opinions which user reveal in their posts to better understand their individual interests. Both the semantic enrichment based on external Web resources and the analysis of users' opinions enable us to generate more valuable user profiles. Given these techniques, we furthermore research the impact of semantic enrichment of microposts on personalization in the microblogging sphere.

As the amount of messages published on microblogging platforms is continuously growing, filtering and retrieval of relevant information (streams) is becoming more and more difficult. Personalized recommender systems [10, 159], which present information tailored to individual users according to their preferences and tastes, allow for supporting users to overcome the information overload problem. In order to deliver personalized recommendations, it is essential to first understand users' information needs and concerns. Research has been done, which is solely based on data from a single source such as Twitter, and focuses on analyzing users' behavior from a single aspect such as the public discussion that users are involved in [93] or the temporal patterns of users' posting behavior [138]. In this thesis, we apply our user modeling framework, which features flexible design choices for constructing user profiles, to conduct large-scale analyses of users' microblogging behavior from different angles across different microblogging platforms and cultural

groups. Such a comparative study based on large microblogging datasets has never been done before and therefore provides unexplored insights for user modeling and personalization based on microblogging data.

Different personalized recommendation systems, which exploit microblogging data for the computation of recommendations, have been developed. For example, personalized recommendations can be computed based on a user' activities on Twitter to rank posts according to the user's preference [44, 108, 148] or suggest to the user interesting information sources to follow [43, 81, 85]. However, there is still a lack of understanding of how different user modeling strategies impact the performance of personalized recommendation systems. Additionally, the real-time nature of information disseminated on microblogging platforms poses new challenges for user modeling and personalization [115, 138]. For example, how do different microblogging-based user modeling strategies influence the performance of personalized recommender systems in the news domain? In this thesis, we evaluate our framework for user modeling based on microblogging data in the context of various personalized recommender systems. We investigate the interplay between trending topics and personal interests to incorporate public trends into the user modeling process and support trend-aware recommendations. Furthermore, we analyze the impact of different design dimensions and design alternatives on the characteristics of user profiles and the performance of recommender systems.

In summary, this thesis contributes to research in the following areas.

Microblogging-based User Modeling Framework. We introduce a framework for modeling users' interests based on microblogging activities and develop a generic software library for generating user interest profiles in various application settings.

Semantic Enrichment for Microblogging-based User Modeling. We exploit different types of resources to enrich the semantics of microblogging activities and analyze the impact of semantic enrichment techniques on the characteristics of user profiles.

Microblogging-based User Modeling for Culture-aware Analytics. Based on our user modeling framework, we analyze user behavior across different microblogging platforms and cultural groups. In addition, we investigate the correlation between our findings and theories about cultural commonalities.

Microblogging-based User Modeling for Personalized Recommendations. We apply our user modeling framework to support various personalized recommender systems and further evaluate the impact of different user modeling strategies on personalization.

1.1 Thesis Outline

This thesis consists of seven chapters. After introducing the motivation of the thesis work in Chapter 1 and the general background and related work in Chapter 2, the main contributions as described above are presented in Chapter 3-6, each of which will start with a motivation of the research questions that are investigated in the corresponding chapter and will conclude with a summary of main findings and contributions.

In **Chapter 2**, we overview related work on user modeling and recommender systems. At the end of Chapter 2, we summarize the key research questions that will be answered in this thesis.

In **Chapter 3**, we introduce *TweetUM* - a user modeling framework that features a variety of user modeling strategies that allow for inferring user's interests and constructing semantically meaningful user profiles based on microblogging data. These user modeling strategies vary in four design dimensions that are described in detail in Section 3.2. In Section 3.3, we present *GeniUS* - a software library, which is implemented based on our user modeling framework, and further demonstrate how this library is able to customize the user modeling process for different application domains. At the end of this chapter, we outline some hypotheses about the impact of different dimensions on the quality of user profiles for further validation.

We exploit two types of resources in **Chapter 4** for the semantic enrichment of microblogging activities: (i) external Web resources that are relevant to microposts and (ii) emotions that are expressed in microposts. We present strategies for linking microposts to external Web resources in Section 4.2 and for identifying emotions in microposts in Section 4.3. We conduct experiments based on data collected from Twitter to evaluate the effectiveness of these strategies and analyze how the exploitation of external Web resources and emotions influence the characteristics of user profiles constructed based on microblogging data.

Utilizing our user modeling framework introduced in Chapter 3 and the semantic enrichment techniques presented in Chapter 4, **Chapter 5** aims at analyzing user behavior across different microblogging platforms. Given various design dimensions featured in our user modeling framework, we compare users' microblogging behavior between two cultural groups (Chinese vs. American users) from different angles. While in Section 5.2 we conduct such a comparative study to reveal the key differences between Chinese and American microblogging practices, Section 5.3 has a focus on examining users' reposting behavior to research the differences in the information propagation patterns on two different microblogging platforms. At the end of both sections, we investigate the correlation between our findings and cultural models from social science research.

In **Chapter 6**, we setup a set of recommendation experiments to evaluate the quality of user modeling strategies provided by our framework. In Section 6.2-6.5, we adapt the process of constructing user profiles for different application settings. In each recommendation experiment, we conduct an in-depth analysis on a large Twitter dataset to understand the influence of different design dimensions and design alternatives on the characteristics of user profiles and further evaluate their impact on the quality of personalized recommendations.

Chapter 7 summarizes our main findings and contributions and answers the research questions raised at the end of Chapter 2. Further, we discuss possible directions for future work.

1.2 Origin of Chapters

Each of the main chapters (Chapter 3-6) is based on at least one peer-reviewed publication, which has been published in conferences related to the research topics of this thesis.

Chapter 3 contains material from two papers that have been published at the *19th International Conference on User Modeling, Adaptation, and Personalization (UMAP'11)*, where it won the best paper award, and at the *2011 Join International Semantic Technology Conference (JIST'11)*. The work used in this chapter originates from Section 3 from each of these papers. Additionally, a short version of the *UMAP'11* paper has been invited to be published at the *23rd International Joint Conference on Artificial Intelligence (IJCAI'13)* in its best papers track.

Chapter 4 contain our work that has been published at the *9th Extended Semantic Web Conference (ESWC'11)*.

Chapter 5 is based on papers published at the *4th International Conferences on Web Science (WebSci'12)* and the *20th International Conference on User Modeling, Adaptation, and Personalization (UMAP'12)*, where it obtained the James Chen best student paper award.

Chapter 6 comprises our findings, which are presented in Section 4 and Section 5 of the *UMAP'11* paper, Section 4-5 of the *JIST'11* paper, and our work that has been published at the *3rd International Conferences on Web Science (WebSci'11)* and the *2011 International Conference on Web Intelligence (WI'11)*.

Chapter 2

Background

In this chapter we introduce background regarding user modeling and recommender systems. We start by giving an overview on the basic concepts and techniques of user modeling. Then we survey related work on constructing user profiles in the Social Web and enhancing the semantics of data in the Social Web for user modeling. We further discuss the state-of-the-art techniques for recommender systems. And last, we summarize the research questions that will be explored in this thesis.

2.1 User Modeling

The term Web 2.0 refers to a new paradigm that was first coined by Tim O'Reilly to address a new generation of Web-based services and tools such as social networking sites, blogs, or wikis [140]. In the era of Web 2.0, people are more involved in publishing and sharing content on the Web. The continuously growing amount of user-generated content on the Web poses new possibilities as well as challenges for understanding users' demands and concerns. In this section we introduce the basic concepts of user modeling and discuss different approaches to user modeling. Based on certain criteria, we present and classify the research efforts on user profile construction in the Social Web. Further, we discuss the Semantic Web technologies that can be used to generate semantically rich information in the Social Web for user modeling.

2.1.1 Overview

User modeling is the process of inferring information about user and representing user information to support a given application [36, 161]. It provides the basis for

a system to adapt to the information needs of individual users. Such adaptation is valuable for various applications such as intelligent tutoring systems [56] that aim to provide customized instruction to students or recommender systems [10] that present tailored information to a particular user based on the user' tastes and preferences. In [95], Jameson et al. identify seven purposes for user modeling including: (i) helping the user find relevant information, (ii) presenting tailored information to the user, (iii) adapting an interface to the user, (iv) providing customized instructions or interventions, (v) giving feedback to the user, (vi) supporting collaboration between users, and (vii) predicting the user's future behavior.

In order to enable the adaptation of systems to different circumstances, user models should be applied. A user model contains the definitions and rules for the interpretation of observations about a user and about the translation of that interpretation into the characteristics in a user profile [89, 99]. The user profile is the data structure that represents a characterization of the user at a particular moment of time [71, 74]. A broad range of user characteristics can be exploited to construct user profiles. For example, in [36], Brusilovsky and Millán summarize five categories of user characteristics including: knowledge, interests, goals, background, and individual traits.

There exist different approaches to user modeling. An overview on user modeling techniques can be found in [39, 99]. In the following, we introduce three types of approaches which have been widely applied in adaptive systems: *stereotyping*, *overlay user modeling* and *user relevance modeling*.

Stereotyping Stereotype user modeling [160], which was developed by Rich and extensively used in early adaptive systems, is one of the oldest user modeling approaches. It tries to categorize all users in a system into several groups, called stereotypes. A user is represented based on her current stereotype that describes specific mixture of characteristics. Then the system only adapt to the user's current stereotype, i.e., all users in the same stereotype are treated in the same way [36]. If the characteristics of a user has changed, a different stereotype can be assigned to the user.

Overlay user modeling An overlay user model represents a user's knowledge, interests, goals, or other features as a subset of domain model, which reflects the expert knowledge of the subject [50, 104]. In an overlay user model, the user is typically characterized in terms of domain concepts and hypotheses regarding the user's knowledge about those concepts. For each concept, the overlay model describes some estimation of the user's knowledge level of that concept.

User relevance modeling The idea of user relevance model is to learn and in-

fer probability that a given concept is relevant for a given user [113, 123]. Therefore, user profiles can be easily represented in a vector space model. The user relevance model is widely applied to personalized information retrieval [35, 102]. For example, by representing both users and documents in the same vector space, similarity measures such as cosine similarity or Jaccard coefficient can be used to estimate whether a given document is relevant to a user [13].

The selection between different user modeling approaches depends on applications where user profiles are used, user characteristics that are exploited to construct user profiles, and other practical as well as theoretical requirements. In the scope of this thesis, we focus on modeling users' preferences based on user and usage data from the Social Web, in particular the microblogging platforms. In the next section, we discuss the state-of-the-art development on user profile construction in the Social Web.

2.1.2 User Profiling in the Social Web

The Social Web is represented by a class of Web-based systems which accomplish an architecture of user participation and collaboration [140]. The value of Social Web is created by the aggregation of many individual user contributions [83]. Social networking sites such as Facebook [61] and Google+ [80] allows users to create networks of friends and share information in their networks. Content-sharing systems such as Delicious [52] and Youtube [186] aim to facilitate the publishing and sharing of user-contributed content. With the advent of microblogging services such as Twitter [178], individuals can post real-time short messages to record thoughts and things that happen in their daily lives. Such short updates can be published using different communication channels (e.g., text messages from mobile phones, text snippets from desktop applications, and share buttons on websites) in various locations.

With the massive amount of information available in the Social Web, there is an urgent need for systems that deliver personalized services, which aim to tailor the information presented to individual users according to the users' demands in terms of content and presentation [71]. There exist various systems that exploit the user information available in the Social Web for personalization such as personalized recommendations [10], personalized information retrieval [74], and personalized navigation [177]. Modeling user and usage information to construct user profiles is crucial for building such personalized systems. In the Social Web, user profiles can be constructed based on different types of user information such as demographic data [47], social network [185], or users' interests that are inferred from

user-generated content [44].

In general, the user profiling process in personalized systems consists of three phases [71, 74]. The first phase is *user information collection*, where information about users is collected using different tools and approaches. The second phase is *user profile construction*, where different modeling approaches and data structures are applied to construct user profiles based on the user information collected in the first phase. The third phase is the *implementation of personalization*, where the constructed user profiles are exploited in order to provide personalized services. The following discussion focuses on the first two phases. In Section 2.2, we will further describe how personalized recommender systems compute recommendations for individual users based on user profiles.

User Information Collection

In this section, we analyze and discuss the research efforts on developing methods for user information collection over two criteria: the *information collection approach* and the *source of information* [74].

Information collection approach Information about users can be obtained in an explicit way where the users need to explicitly provide information to the system or in an implicit manner where the information is gathers without any effort from the users. In the explicit approach, a user can supply information to a system by specifying her interests for items (e.g., movies [171], music [141], news articles [148], etc.), or by giving positive or negative relevance feedback about the information delivered by the system [22]. For example, Carmagnola et al. present a system called iCity that exploits users' social tagging activities in context of cultural heritage domain to construct and update user profiles [38]. The system then recommends cultural events taking place in a city according to individual users' interests represented in the user profiles. Hannon et al. collect Twitter messages that are published by a user and other users that she follows to model her interests using a bag-of-words approach [85]. One problem with the explicit approach is that people may not be willing to provide personal information such as demographic data or personal interests due to privacy concerns [100]. In comparison to the explicit approach, the main advantage of implicit user information collection is that it does not place any burden on users for generating user profiles [71]. The implicit approach aims to automatically collect user information by analyzing log data such as queries submitted by a user, utilizing information from a user's interaction with a system, or processing any stored content to infer individual users' interests [74]. For example, different from the work

conducted by Hannon et al. [85] which only utilizes explicit information (textual content of tweets), some research focuses on extracting implicit information from microblogging activities such as political preferences of individual users [78, 176], emotions that are expressed in microposts [51], or latent topics that are inferred from a collection of microposts [152].

Source of information The user information collection also varies depending on the sources where the information is obtained. Some systems only gather information from single Social Web applications such as ratings from a movie recommender system [171], users' tagging behavior [38], or microblogging activities from Twitter [184]. The advantage of collecting information from a single source is that the user information is represented in a consistent format. However, it is not capable of capturing the user information distributed in various Social Web applications, which can be beneficial for supporting personalized systems. For example, the integration of user information from multiple systems help recommender system deal with spam and cold start problems [128]. Abel et al. present an approach for user modeling across Social Web systems [8]. They present strategies that allow for the aggregation of profile information including demographic information (e.g., name, location, etc.) and tag-based profiles distributed in different Social Web systems such as Facebook, Delicious, and Twitter. The aggregated user profiles reveal more facets about individual users. Furthermore, the authors investigate the impact of aggregated profiles on personalization and discover that the aggregated profiles improve the performance of tag recommender systems significantly.

User Profile Construction

The second phase focuses on user profile construction based on the information collected from the first phase. The discussion presented in the following is carried out over five criteria including the *user features* that are exploited, the *scope of interests*, the *user profile representation*, the *dynamism of user profile*, and whether the *semantics* of user information is inferred for constructing user profiles [71, 74, 175].

User features The user features, which are exploited to construct user profiles, vary depending on the personalization functionality and the Social Web systems where user information is collected. For example, a variety of user features can be used to construct users profiles based on Twitter activities. In [47], Cheong et al. collect demographic information of Twitter users such as clients and devices that are used to post messages, and gender information which is

explicitly claimed by the users or inferred based on the writing styles or profile images. The demographic information is then used to analyze the characteristics of users who contribute to the discussion of a trending topic. Hecht et al. propose a machine learning approach for identifying a user's location based on her Twitter messages [88]. The exploitation of geographic information allows for enabling location-based personalized services such as recommending points of interests (POIs) in a city [7]. Recently, researchers have also started investigating methods to extract sentiments that are expressed in microposts [76, 167]. While the features discussed above are extracted based on Twitter activities themselves, social network is a feature that explores the social relationships (followee and follower) of Twitter users and can be used to model individual users' interests [44] or identify influential users in Twitter network [110]. Additionally, researchers investigate theories and methods from other disciplines (e.g., social science, psychology, etc.) to extract user features based on microblogging data such as personality [79], learning style [86], cultural commonality [70], and political preference [78].

Scope of interests In many personalized system, user profiles are created to represent individual users' interests based on which personalization is provided. The users' interests can be categorized into long-term or short-term interests [74]. While the long-term interests exhibit persistent interests of individual users, the short-term interests are ephemeral interests that usually reflect the users' information needs during a short period of time. Huang et al. apply statistical approaches to explore the temporal patterns of hashtags in Twitter [93]. They use standard deviation to measure the spread of a hashtag in Twitter network, representing how long a hashtag remains in use. Their study reveals the phenomenon of micro-memes, where the hashtags created for emergent topics are used widely for a few days and then die-out quickly. Therefore, hashtags can be utilized to model a user's short-term interests for supporting personalization such as recommending trending events that are related to certain hashtags [107]. In addition, the long-term and short-term interests can be integrated to provide a comprehensive understanding of users' demands and concerns [49, 117]. Li et al. propose a method to integrate the long-term and short-term online reading preferences of individual users to recommend news articles [114]. While the long-term user profile of a user is constructed using a time-sensitive weighting scheme [57] based on the user's entire history, the short-term profile of that user is constructed by analyzing her latest activities.

User profile representation Several techniques and data structures can be applied to construct user profiles in personalized systems. Following the classifica-

tion reported in [65, 74], here we discuss two different types of user profile representations: *vector-* and *semantic network-based* user profiles.

- A *vector-based user profile* is represented using a vector of terms and associated weights. The weights are computed by a certain term weighting scheme such as TF, $TF \times IDF$, or time-sensitive weighting scheme [57]. In vector-based user profiles, the terms can be represented by words or concepts that are extracted from user-generated content. For example, Hannnon et al. construct user profiles based on Twitter messages and represent user profiles in vector space model [85], where the terms are represented by words extracted from Twitter messages and the associated weights are computed using a term frequency-based weighting scheme. Alternatively, the terms can be represented by semantic concepts such as named entities that are extracted from textual content [72]. Additionally, a user can have more than one user profile represented in multiple vectors [74]. For example, Li et al. use one vector to represent the short-term interest profile of a user and another vector to represent the long-term interest profile of the same user [114].
- In a *semantic network-based profile*, the user's interests are modeled in a network structure of terms and related terms [74]. Weights can be assigned to the terms and their related terms, and the links between them. In comparison of vector-based user profiles, semantic network-based profiles allow for describing the relationships between a term and its associated terms. Such relationships can be derived using existing thesauruses such as WordNet [62] or external knowledge sources such as DBpedia [12]. In InfoWeb [73], a personalized information filtering system for online digital documents, semantic network-based user profiles are applied to model user interests. Initially, each user profile is made up of a set of concepts that represent a user's interests. As the user continuously interacts with the system, her user profile is updated by adding more concepts to the semantic network and links between the concepts.

Dynamics of user profile Information stored in static user profiles is less likely to change over time. Such information can be, for example, personal background, personality, or demographic information, and is not subject to continues updates [74]. Information stored in dynamic user profiles, on the other hand, evolves over time. For example, Gentile et al. propose an approach to dynamically model user expertise based on information communication exchange such as emails [72]. In contrast, user profiles that describe short-term user interests are usually updated frequently over time [73, 114].

Semantic To construct user profiles, user information can be collected from various Social Web systems. However, the lack of interoperability between different systems makes the reuse and interlinking of user information difficult [30]. To overcome this problem, the Semantic Web technologies, which provide common standards to model information on the Web, can be applied to make information across various Social Web systems interoperable [175]. The research efforts on integrating the Semantic Web with the Social Web can be distinguished between the ones that directly apply the Semantic Web technologies to build Social Web applications and the ones that focus on extracting semantically rich data from existing information in the Social Web [30]. In Section 2.1.3, we will further discuss how to enhance the semantics of information in the Social Web for user modeling.

2.1.3 User Modeling for the Social Semantic Web

The Semantic Web is "not a separate Web but an extension of the current Web in which information is given well-defined meaning, better enabling computers and people to work in cooperation" [21]. While the Social Web is designed to facilitate user participation, the Semantic Web aims to define extensible standards for information exchange and interoperability so that data can be shared and reused across applications [90]. Berners-Lee describes a layered Semantic Web architecture which consists of a set of standards [19]. The Resource Description Framework (RDF) [179] is used to describe resources and relationships between those resources. In RDF, two resources and a binary relationship between these resources is called a subject-predicate-object triple, or a RDF statement, which describe the property (predicate) of a resource (subject) with some value (object). To make the data represented in RDF interchangeable across applications, a set of RDF statements can be serialized in different formats such as RDF/XML [15], N-triples [16], or Notation3 [20]. The RDF schema specifies a set of classes and properties to describe ontologies [31]. Web Ontology Language (OWL) can be used to model advanced axioms such as symmetric or transitive properties in ontologies [127]. Moreover, RDF data can be stored in RDF repositories [34] and queried using query language such as SPARQL [149].

The integration of Social Web and Semantic Web is leading to the "Social Semantic Web" which describes a network of interlinked and semantically rich data [30, 135]. It brings Social Web with knowledge representation languages and formats from the Semantic Web. With the Semantic Web technologies, information in the Social Web can be represented using common shared models such as ontologies and therefore can be reused and shared across applications.

According to the definition given by Gruber, an ontology is an explicit specification of the conceptualization of a domain [82]. In the Social Web, ontologies can provide shared and uniform models to represent different artifacts in the Social Web such as people, documents, and tags. In [175], Torre gives an overview on ontologies and vocabularies for user modeling in the Social Web such as *FOAF*, *SIOC*, and ontogigies for modeling tagging activities.

FOAF The *Friend-of-A-Friend* (FOAF) ontology specifies a set of classes and properties to describe people as well as the relationships between people [32]. For example, the *foaf:Person* and *foaf:Documents* classes are used to describe people and the documents that people create. Individuals can also apply the *foaf:knows* property to create social networks by specifying their connections to people that they know. The FOFA files can be exported directly in some Social Web applications such as LiveJournal (a blogging community site) or via third-party components for social networking sites such as Facebook [164]. Given the shared knowledge representation, multiple FOAF files, which are distributed in the Web, can be combined to provide an aggregated view of the network across various systems.

SIOC While FOAF aims to model people and their networks, *Semantically Interlinked Online Communities* (SIOC) project provides a lightweight ontology for describing the structure of online communities as well as user-generated content in the Social Web such as blog posts and topic threads in online discussion forums [27]. The SIOC core ontology, which consists of a set of RDF classes and properties, allows for interlinking information across community sites using RDF data and can also be combined with other existing ontologies such as FOAF. Bojars et al. describe the use of SIOC, FOAF, and other vocabularies for interlinking and reusing user data across various social applications [25]. Passant et al. apply the SIOC ontology together with other domain ontologies to enrich the semantics of blog posts and further present experimental results that show the semantic enrichment of blog posts improves the search experience in comparison to free-tagging approaches [144].

Tag ontologies Social tagging describes the process by which a group of users assign unstructured keywords (tags) to online resources. Due to the lack of predefined taxonomies, social tagging systems rely on "shared and emergent social behaviors" [122]. The term *folksonomy* depicts the structures that emerge from social tagging systems [126]. Mathes discusses the limitations of tagging and lists two major problems with folksonomy systems [124]. Firstly, tags have little semantics, which makes it difficult to consistently represent a user's interests. Secondly, it's difficult to aggregate tagging data from different systems since most tagging systems interpret the meaning of tags in their

own specific ways. To overcome these limitations, tag ontologies has been developed and applied to provide a uniform structure and semantic representation in folksonomy systems. Mika investigates how to define ontologies to materialize the emergent semantics of follsonomies [134]. Newman et al. introduce the *Tag Ontology* which allows for specifying the relationship between an user, a resource and one or more tags [139]. The Tag Ontology has been applied in systems such as Revyu.com [87], a social tagging systems for sharing reviews. The *Meaning of a Tag* (MOAT) ontology aims to provide a meaning for free-text tagging through semantic annotation [145]. It provides a framework that allows people to annotate the content by selecting appropriate URIs or using resources from existing knowledge bases such as DBPedia [12]. For example, Abel et al. use the *moat:meaning* property to unambiguously describe the meaning of a tag in a given context [1]. For a comprehensive overview and comparison of tag ontologies, we refer the reader to [98].

To leverage the wisdom of the crowds in the Social Web to generate semantically rich data, some research efforts focus on applying the Semantic Web technologies to build various Social Web applications such as webblog [41, 97], wikis [105, 172], and social bookmarking systems [136, 183]. For example, Revyu.com is an online service where users can create reviews for items such as restaurants, books, and movies [87]. It combines some features of the Social Web applications such as tagging with the Semantic Web standards to build a review website. Each review is modeled in RDF and can be queried via a SPARQL endpoint. The *Semantic MicrOBlogging* (SMOB) is a service that allows for the generation of semantically rich microblog posts, which can be propagated through microblogging services like Twitter [146]. SMOB uses existing ontologies such as FOAF and SIOC to represent the users, their properties, and service information.

While the research discussed above applies the Semantic Web technologies to directly create data that can be consumed in various applications in the Social Web, some efforts aim to infer semantics from existing social data such as the microposts that people have already published. Rowe and Stankovic present an approach for the semantic enrichment of Twitter activities by extracting DBpedia concepts from Twitter messages [165]. Individual Twitter activities are modeled in a semantically rich and structured format and can be further woven into the Web of Linked Data [23]. The Linked Data project proposes four basic design dimensions as follows to publish, share and connect pieces of data on the Web using the Semantic Web standards.

1. Use URIs as names for things.

2. Use HTTP URIs so that people can look up those names.

3. When someone looks up a URI, provide useful information, using the standards (RDF, SPARQL).

4. Include links to other URIs so that they can discover more things.

In [165], Rowe and Stankovic apply these principles to enable the Twitter activities to be published as linked data. The DBPedia concepts that are related to a Twitter message as well as the medata of that message, such as the creation time and the user information, are defined via resolvable HTTP URIs which can be looked up using SPARQL. Furthermore, the authors discover that the semantic enrichment of tweets using external knowledge sources such as DBPedia on the Web of Linked Data is beneficial for supporting the alignment of events with Twitter messages.

2.2 Recommender Systems

In this section, we first define the general task of recommender systems. Then we introduce the two most popular approaches for the computation of recommendations.

2.2.1 Overview

Recommender systems are software tools and techniques which provide suitable recommendations for items to individual users [37, 157]. A recommender system normally focuses on a specific type of item such as books, musics, or news. In order to compute recommendations, three types of information are exploited [159]: (i) items that are available for recommendations and properties that describe the items, (ii) users of recommender systems and information about the users, (iii) transactions that describe relations between users and items.

Formally, recommendation task can be formulated as follows [10]: Let U be the set of all users in a system and Let S be the set of all possible items to be recommended in the system. A utility function f is used to measure the usefulness of an item $s \in S$ to a user $u \in U$. The function f is defined as $f : U \times S \rightarrow R$, where R is an ordered set that is made up of nonnegative integers or real numbers within a certain range. For each user $u \in U$, the task of recommendation is to choose such item $s' \in S$ that maximizes the user's utility.

The spaces of both items and users can be very large, ranging in hundreds of thousands or even millions in some systems [116]. For each user $u \in U$, the user

profile can be constructed by exploiting user characteristics (e.g., interests, demographic information, knowledge levels, etc.). Similarly, each item $s \in S$ is also represented by a set of properties.

Utility can be represented by various functions. For example, in MovieLens project [171], which focuses on building movie recommendation applications, utility is represented by ratings of movies. Initially, a user rates the movies that she has already seen on a scale of 1 to 5. The goal of recommender system is to estimate the ratings of the movies that are not rated yet by the user and generate recommendations based on the estimated ratings. In context of tweet recommendations, which recommend Twitter messages to individual users, the utility of an item (Twitter message) for a user can be represented by a binary rating which indicates whether the user is interested in that message [44].

Various recommendation techniques have been developed throughout the last two decades. In general, three categories of recommendation approaches can be distinguished [10].

- *Collaborative filtering-based approach* that recommends to a user items that other users with similar tastes liked in the past;
- *Content-based approach* the recommend items to a user by finding items similar to the ones that the user liked in the past;
- *Hybrid approach* that combines collaborative filtering and content-based methods.

In the next sections, we describe the collaborative filtering and content-based approaches in detail. We refer the reader to [37] for details of the hybrid approach.

2.2.2 Collaborative Filtering Recommender Systems

Collaborative filtering recommender systems try to recommend items to a particular user based on the items that has been previously rated or seen by other users [54, 171]. Therefore, collaborative filtering techniques compute recommendations based on the user profiles of other users who have similar tastes and preferences. More formally, collaborative filtering-based approach estimates the utility $f(u,s)$ of item s for user u based on the utilities $f(u_j,s)$ that have been assigned to item s by those similar users $u_j \in U$. The recommendation process is based on a so-called *user-item matrix* which consists of all users, items and the users' existing ratings for items. In general, there exist two classes of methods for collaborative filtering: *memory-based* and *model-based* methods [10, 53].

In memory-based collaborative filtering systems, the entire user-item matrix, which stores all the existing user-item ratings, is directly used to estimate the ratings for new items. In general, there are two ways to compute recommendations in memory-based systems: *user-based* and *item-based* filtering. User-based filtering first matches the user profile of a user against the user profiles of other users in the system to identify a set of users (neighbors) who have similar preferences. Then the interest of that user for an new item is evaluated by aggregating the ratings given by the top-k most similar users for the same item [54]. In user-based systems, user profiles are usually represented as vectors. Then the similarities between user profiles can be measured using metrics such as cosine similarity or Pearson correlation coefficient [158]. These metrics can also be used to compute the similarities between items. Item-based filtering estimates the interest of a user for a new item based on the ratings of the most similar items in the system [53, 171].

Model-based recommendation algorithms apply machine learning techniques to learn a predictive model based on a user-item matrix. The goal is to identify latent factors which are used to model the user-item interactions in a system. The model is trained using existing data and then applied to compute recommendations. Bresse et al. investigate two probabilistic approaches for leaning the model: clustering and Bayesian netowrk [29]. Another group of model-based algorithms, which becomes popular through the Netflix competition [103], is based on matrix factorization techniques such as Singular Value Decomposition (SVD) [17] and Latent Dirchlet Allocation (LDA) [24].

In contrast to content-based systems which focus on analyzing content of items for recommendations, collaborative filtering recommender systems aim to exploit other users' ratings to compute recommendations. Therefore, collaborative filtering techniques are capable of dealing with any kind of items, even the ones that are not similar to those which have been rated in the past [10]. Nonetheless, collaborative filtering recommender systems have their own limitations that are summarized as follows.

New user problem Collaborative filtering systems suffer from the new user problem, i.e., the systems would not be able to learn the preferences of a user and make accurate recommendations until the user gives a substantial number of ratings. Several recommendation systems employ hybrid approach, which combines collaborative filtering and content-based techniques, to address this problem [37, 173].

New item problem Since collaborative filtering methods rely on using other users' activities to estimate the interest of a given user for an item. Therefore, the item must have been rated or seen by other users in order to compute recom-

mendations. Moreover, in many collaborative filtering systems, most users only interact with a very small fraction of all items, which makes the user-item matrices immensely sparse. Due to the lack of available information such as users' ratings the quality of recommendations may not be satisfying. In contrast, content-based systems are better able to cope with the new item problem.

2.2.3 Content-based Recommender Systems

Content-based recommender systems, as the name implies, rely on the content of items in the systems such as documents that contain textual information, online resources that are tagged with keywords, or movies that are described with genres, actors, subjects, etc. In content-based approach, the utility of an item for a user is usually computed based on the ratings that have been assigned by the same user to similar items.

The recommendation process consists of three steps [119]. The fist step focuses on analyzing the content of items to extract relevant structured information for the next steps. The main responsibility of the second step is to construct *item profiles*, which exploit a set of properties to characterize items, as well as *user profiles* that describe users' tastes, preferences, and information needs. Finally, the recommender tries to find relevant items for a user by matching her user profile against the profiles of items to be recommended.

In content-based recommender systems, items are often represented by textual features such as keywords that are extracted from various types of content (e.g., Web pages, news articles, descriptions of movies, etc.). As the content-based approach has its roots in information filtering and information retrieval research [13], most content-based systems use retrieval models such as the vector space model to construct item profiles as well as user profiles [119]. For example, the *item profile* of item s, denoted as $ItemProfile(s)$, can be represented by a vector in a multi-dimensional space, where each dimension corresponds to a keyword. And various weighting schemes such as TF or $TF \times IDF$ can be applied to determine the weight of each element in the item profile [10].

Given a user u, her user profile $UserProfile(u)$ is generated by analyzing the content of the items that user u has already rated or seen in the past and is usually represented by keywords or semantic concepts extracted from the content of those items. Similar to item profiles, user profiles can also be represented in the vector space mode with a variety of weighting schemes [71]. In addition, machine learning techniques can be applie to learn and update user profiles [147].

Given the vector representation of $UserProfile(u)$ and $ItemProfile(s)$, denoted as $\vec{p}_u$ and $\vec{p}_s$ respectively, the utility $f(u,s)$ of item s for user u can be computed using similarity measures such as the cosine similarity [13]. The recommender system then generates the recommendations for the user by ranking the candidate items based on their utilities.

The content-based method has several advantages described as follows [119].

User independence In comparison to collaborative filtering techniques, content-based recommender systems exploit solely the history of a user to construct the user profile for the computation of recommendations. Therefore, the algorithm does not require any extra information from other users.

Transparency Since both item and user profiles are constructed with features that are extracted from the content, a content-based recommender system allows for providing explanations on how the system works by describing explicitly the features that cause a particular recommendation. In contrast, the only explanation that can be provided for an recommendation based on collaborative filtering is that some (unknown) users with similar preferences liked that item [14]. The explicit explanations can help users judge whether they should trust the recommendations [48, 174]. For example, Cramer et al. conduct a user study to investigate the impact of transparency on user trust in content-based recommender systems [48]. They discover that providing explicit explanations to users increases their acceptance of the systems.

New item In content-based recommender systems, it is possible to recommend items that are not yet rated by any user. Therefore, the systems do not suffer from the new item problem. The content-based techniques can be applied to recommend emerging items such as Twitter messages related to breaking news [148]. In contrast, in collaborative systems, new items need to be rated by a substantial number of users in order to generate accurate recommendations.

However, content-based approach has its own limitations that are summarized as follows [10, 14].

Limited content analysis Content-based recommender systems are limited by the number and type of features that are used to represent the items to be recommended. Therefore, content-based systems need to first extract features from the content of items to construct item and user profiles. In many applications, the feature extraction requires domain knowledge or ontologies [133].

Overspecialization For each user, content-based approach tries to identify the most similar items based on the user profile to compute recommendations. This results in a lack of serendipity, i.e., the recommendations may have a limited degree of novelty.

New user In order to understand users' preferences and deliver accurate recommendations, a content-based recommender needs to collect sufficient number of ratings for each user in the system. As a consequence, for a (new) user who only has few ratings or no rating at all, the system is not capable of constructing the user profile and further providing reliable recommendations.

2.3 Research Challenges tackled in this Thesis

In this chapter, we discussed the general background of user modeling and recommender systems in the Social Web. The rest of this thesis will focus on researching user modeling based on microblogging data and applying user modeling techniques to support various applications such as personalized recommender systems. The research challenges, which will be tackled in this thesis, are summarized as follows.

Microblogging-based User Modeling Framework. Recently, researchers started to exploit microblogging activities to infer specific attributes of a user such as the user's location [46], political orientation [78], or influential power [42]. Yet, there exists no generic user modeling framework for inferring users' interests from microblogging data and supporting personalization in different contexts.

- How can users' personal interests be inferred from microblogging activities?
- How can we generate semantically meaningful user interest profiles that can be applied in different application domains?

In Chapter 3 we will answer these questions and introduce a user modeling framework that features different design dimensions and design alternatives for constructing semantically meaningful user profiles based on microblogging activities. Further, we will present a software library that allows for the generation of customized user profiles for a particular application setting.

Semantic Enrichment for Microblogging-based User Modeling. In Section 2.1, we surveyed various approaches for generating semantically rich data in the Social Web [30, 83, 175]. However, little research has been done to understand the semantics of individual microblogging activities for user modeling.

- How can we enrich the semantics of individual microblogging activities?
- How does the semantic enrichment impact the characteristics and quality of microblogging-based user profiles?

Answers to these questions will be presented in Chapter 4 where we investigate different methods for enriching the semantics of microposts and analyze their influence on the characteristics of constructed user profiles.

Microblogging-based User Modeling for Culture-aware Analytics. Initial work has been done to investigate a variety of user characteristics to understand users' microblogging behavior [96, 120]. Most research focused on analyzing user behavior from single aspects such the usage of hashtags [93] or the evolution of users' interests over time [138]. Moreover, there exists little research on studying the user behavior across different cultural groups.

- How does the microblogging behavior vary between different cultural groups?
- Do differences in users microblogging behavior correlate with cultural theories in social sciences?

These questions will be answered in Chapter 5 by applying our user modeling framework to compare users' microblogging behavior on two microblogging platforms from different angles and by interpreting our findings with theories about cultural commonalities in social science research.

Microblogging-based User Modeling for Personalized Recommendations. The huge amount of microposts posted every day makes the retrieval of relevant information more and more challenging. Researchers have developed applications which aim at understanding users' preferences and providing personalized services to individual users based on microblogging data [44, 85, 108]. However, the impact of different design dimensions and design alternatives on personalization in the microblogging sphere has not been studied extensively yet.

- How do the different user modeling strategies influence the quality of user profiles and the performance of personalized recommender systems?
- What is the impact of incorporating trends and domain-specific knowledge into the user modeling process on the quality of personalized recommender systems?

We will answer these questions in Chapter 6 by experimenting and evaluating different user modeling strategies in various personalized recommender systems. We will investigate and evaluate methods to integrate public trends into user profiles for supporting trend-aware recommendations.

Chapter 3

Microblogging-based User Modeling Framework

After the previous chapter has presented background knowledge on user modeling, in this chapter we introduce a user modeling framework where user interests are derived from microblogging activities. The framework builds the basis for various applications such as microblogging behavior analytics (see Chapter 5) or personalized recommender systems (see Chapter 6). The main contributions of this chapter have been published in [5, 66].

3.1 Introduction

On microblogging platforms such as Twitter, people publish short messages to share their thoughts and things that happen in their daily lives. The plethora of digital traces, which people leave in the microblogging sphere, provides possibilities for modeling user preferences and delivering personalized services. In comparison to other Social Web services like Last.fm, which allows for the deduction of users' musical taste [63], or Flickr, which primarily provides information to infer users' interests in locations or events [153], microposts on Twitter are not restricted to a certain domain. Instead, users can discuss about any topic they are interested in or concerned with which makes it worthwhile to explore microblogging activities for supporting valuable external applications. Sakaki et al. developed an early warning system that enables prompt reporting of earthquakes by collecting and analyzing Twitter messages containing relevant keywords such as "earthquake" or "shaking" [168]. Mathioudakis et al. introduce a system called *Twittermonitor* that allows for detecting trending topics which are represented by named entities or bursty key-

words identified from Twitter streams [125]. These applications mainly utilize the wisdom of the crowds as a source of information rather than relying on individual microposts and individual user behavior.

Understanding individual microblogging activities and individual behavioral patterns can be considered important for better supporting applications that aim for personalization. For example, given the huge amount of information disseminated daily on Twitter, user profiling that supports users in ranking sources to follow [85] or selecting content to read [44] is becoming crucial. Recently, researchers started to exploit microblogging activities to understand users preferences and behavioral patterns. Cheng et al. investigate how to infer a user's location based on the content of tweets [46]. Golbeck et al. present a method to measure users' political orientations [78]. In [42], the authors study the dynamics of user influence across topics and time. Yet, little research has been done that focuses on understanding the semantics of individual microblogging activities and inferring user interests from these activities. Making sense of individual activities for user modeling and personalization is—due to the shortness of microposts—a non-trivial problem that we investigate in this thesis.

In this chapter, we introduce a framework for generating users' interest profiles from microblogging activities. A key challenge that we deal with is the generation of semantically meaningful user profiles from microblogging streams which can be consumed by different applications. Laniado and Mika analyze the semantics of hashtags, words that start with "#", and propose metrics that characterize hashtags as descriptors for retrieving information in Twitter [109]. Chen et al. exploit the social network of a user as well as the general popularity of the URLs in Twitter to model user preference for recommender systems [44]. However Chen et al. do not investigate user modeling in detail, but represent Twitter messages of a user by means of a bag of words. Neither hashtag-based nor bag-of-words representation explicitly specify the semantics of microposts. Rowe et al. propose the use of contextual information to enrich the semantics of tweets [166]. The authors also mention user profiling as one of the applications that might benefit from such semantics, but do not further investigate user modeling in the microblogging sphere. To close this gap, our user modeling framework that is presented in this chapter leverages microblogging activities for constructing user profiles based on the semantics extracted from the microposts.

Exploiting the microblogging streams promises to be of benefit for applications that need to understand the demands and concerns of the people. Different applications may have specific demands for user profiles. For example, an online book store that features book recommendation functionality requires information about a user's interests in books, a music recommendation platform needs to gather informa-

tion about a user's musical taste, and a movie recommendation system has to infer a user's preferences in movies. Those applications thus require user profiles that represent domain-specific interests for individual users. To fulfill the demands of different applications, our user modeling framework offers flexible design choices that allow for generating customized user profiles for various applications.

In this chapter, we will answer the following research questions.

- How can users' personal interests be inferred from microblogging activities?
- How can we generate semantically meaningful user interest profiles that can be applied in different applications?

In Section 3.2, we will introduce a microblogging-based user modeling framework for inferring users' interests and describe the design space of user modeling strategies. In Section 3.3, we will present a generic software library implemented based on our user modeling framework which allows for the generation of semantically meaningful user profiles to support various applications. We conclude this chapter with a short discussion and a list of hypotheses that will be further investigated in the subsequent chapters.

3.2 TweetUM - Tweet-based User Modeling Framework

The user modeling strategies that are proposed and discussed in this chapter aim to generate user profiles that reflect the interests of a user. Hence, the user profiles will describe to what extent a user is interested in a certain topic. The generic model that can thus be applied for representing user interests can be specified as follows (cf. Abel et al. [8]).

Definition 1 *The profile of a user $u \in U$ at a given timestamp $time$ is a set of weighted topics where with respect to the given user u for each topic $c \in C$ its weight $w(u,c,time)$ is computed by a certain function w.*

$$P(u,time) = \{(c, w(u,c,time)) | \forall c \in C\} \tag{3.1}$$

Here, U denotes the set of users while C denotes the set of concepts used to represent the topics of interests. In addition to the model utilized by Abel et al. [8], we make the weighting function $w(u,c,time)$ time-aware, i.e. the interest scores depend on the time frames for which the profile is requested. To facilitate the interpretation and

design dimension	**design alternatives** (discussed in this chapter)
topic modeling	(i) hashtag-based, (ii) category-based, (iii) entity-based, or (iv) LDA-based
enrichment	(i) micropost-only-based enrichment or (ii) linkage and exploitation of external Web resources (propagating topics)
temporal constraints	(i) specific time period(s), (ii) temporal patterns (*weekend*, *night*, etc.), or (iii) no constraints
weighting scheme	(i) TF, (ii) TFxIDF, (iii) time-sensitive TF, or (iv) time-sensitive TFxIDF

Table 3.1: Design space of Twitter-based user modeling strategies.

processing of such user profiles, we typically normalize user profiles to make the sum of all weights in a profile equal to 1: $\sum_{c_i \in C} w(u, c_i, time) = 1$. With $\vec{p}(u, time)$ we refer to $P(u, time)$ in its vector space model representation, where the value of the i-th dimension refers to $w(u, c_i, time)$.

When designing a user modeling strategy that generates user profiles according to the above definition, there are a couple challenges and design decisions that have to be tackled. We list them in Table 3.1.

Topic modeling. What are the actual topics of interests? According to Definition 1, one has to decide what kind of concepts $c \in C$ are used to model topics. As we aim to obtain an understanding of a user's interests so that we can provide, for example, news recommendations that adapt to these interests, we have to investigate what kind of concepts are best suited to represent a user's topics of interests. In particular, following Table 3.1 we propose four approaches to constructing profiles that differ with respect to the type of topics in C: hashtag-based, category-based, entity-based and LDA-based topic modeling. We discuss them in detail in Section 3.2.1.

Enrichment. The user modeling strategies that we employ in this thesis exploit Twitter messages posted by a user u to construct the corresponding profile $P(u)$. A core question that we investigate in this thesis is whether these short messages are a sufficient basis for building user interest profiles that can be applied to provide personalization (see *tweet-only-based enrichment*, Table 3.1) or whether further enrichment is beneficial to the Twitter-based user modeling (see *linkage and exploitation of external Web resources*, Table 3.1). In particular, we explore whether enrichment with topics extracted from Web resources such as online news articles that are linked from the micropost adds value to the user modeling. Here, we investigate strategies that exploit URLs

which the users explicitly posted in their tweets as well as strategies that also aim to link tweets which do not contain a URL with related Web resources. Additionally, while the content of microblogging posts can be objective in nature, it can also reflect different opinions of individuals such as positive and negative emotions [51]. Therefore, we further enrich the semantics of microblogging activities by exploiting the emotional states that are expressed in the microposts (see Section 3.2.2).

Temporal constraints. A third dimension that we investigate in the context of microblogging-based user modeling is given by *temporal constraints* that are considered when constructing the profiles (see Table 3.1). For example, is it useful to exploit the entire history of a user's Twitter timeline when constructing her user interest profile? To answer this question, we first study the nature of user profiles created within specific time periods. Second, we examine certain time frames for creating the profiles (see Section 3.2.3).

Weighting scheme. Given a topic modeling strategy, an enrichment strategy and a strategy for incorporating temporal constraints, one has to select an appropriate weighting scheme $w(u,c,time)$ that assigns a weight to each topic of interest. The weight specifies to what extent a user is interested into the topic c. Our framework provides different weighting schemes ranging from term frequency based methods that count the number of occurrences of c in u's tweets (see TF, Table 3.1) to more advanced time-sensitive weighting schemes that also incorporate a temporal decay (see Section 3.2.4).

Together, the different user modeling building blocks form a rich microblogging-based user modeling framework. By selecting and combining the different design dimensions and design alternatives, we obtain a variety of different user modeling strategies that will be analyzed and evaluated in this thesis.

3.2.1 Topic Modeling

When building a user profile according to Definition 1 that reflects the interests of a user, a core design decision is the selection of appropriate concepts $c \in C$ that specify into *what* a user is interested. A naive strategy would be to represent the topic of interests C via the terms that a user is mentioning in her tweets, i.e. a so-called *bag-of-words* strategy. Our microblogging-based user modeling framework goes beyond bag-of-words and provides four main strategies for modeling a user's topics of interests: hashtag-based, category-based, entity-based and LDA-based topic modeling. Below, we describe these strategies in more detail.

id	content	publish date
t1	I am happy that francesca schiavone is becoming #sport idol of this year http://bit.ly/grIlmM	9:10AM Fri. Dec. 3, 2010
t2	Just got a new iPad lol #excited	8:34PM Sat. Dec. 25, 2010
t3	RT @cnnsportsnews: Fed Cup champions Italy held on opening day of 2011 campaign http://bit.ly/WwSwou	4:02PM Fri. Feb. 4, 2011
t4	Awesome, love the new Garageband for iPad http://bit.ly/fkj3YM #apple	8:55PM Sat. Feb. 27, 2011

Table 3.2: Example microposts of a user

hashtag-based	category-based	entity-based
sport	Sports	Francesca Schiavone
excited	Technology_Internet	iPad
apple		Fed Cup
		Italy
		Garageband
		Apple

Table 3.3: Topics of interests extracted from the example microposts in Table 3.2

Hashtag-based topic modeling utilizes hashtags to represent individual user interests. Hashtags (e.g., #web, #technology, #umap2012) are often used as identifiers in Twitter messages to join certain discussions on Twitter [93]. Using a hashtag as filter, people can monitor discussions that refer to the corresponding topic. Table 3.2, lists example microposts published by a user during a certain period of time. The user's topics of interests are modeled using the hashtags (e.g., #sport, #excited, #apple) that are extracted from the textual content of microposts (see Table 3.3). However, the semantic meaning of hashtags might depend on the usage context within Twitter discussions. For example, without any further contextual information, the meaning of the hashtag #apple is ambiguous as it could refer to a kind of fruit or the name of a company. Therefore, more advanced topic modeling strategies are needed so that external applications such as personalized recommender system can better understand the semantic meaning of the topics of interests.

To provide richer contextual information for modeling topics of tweets, we further examine strategies that extract categories (e.g., politics, sports, eduction) and entities (e.g., persons, organizations, events) from Twitter messages and related Web resources. The *category-based topic modeling* classifies microposts into broad categories using existing services. For example, we differentiate between 18 broad

categories proposed by the OpenCalais service[1] which is one of the services that is utilized by our semantic enrichment component (see Secion 3.3). Categories are more abstract and broader than hashtags. Therefore, *category-based topic modeling* could be beneficial for applications such as news readers that may need to summarize content. In addition, categories are more stable than hashtags. While hashtags are very often used for trending discussions or single events, categories might live for much longer period of time. The *entity-based topic modeling* strategy exploits entities which—in comparison to categories—refer to more concrete topics, for example, instances of the class *person* such as "Barack Obama" or concrete *locations* such as "London".

Table 3.3 presents, for the user whose microposts are listed in Table 3.2, her topics of interests that are represented using categories and entities respectively. Both categories and entities are identified via URIs so that the semantic meaning of category-based and entity-based profiles are explicitly defined. For example, by dereferencing the unique URI[2] that are associated with the entity *Apple*, the entity, which refers to the name of a company, is semantically disambiguated. With the semantically meaningful topics, our user modeling framework is able to construct user profiles to serve different applications. Furthermore, given the four example microposts listed in Table 3.2, *entity-based topic modeling* is able to produce five distinct topics, which are more than the amount of hashtags or categories that are extracted from the textual content. While the category-based profiles give a summarization of the topics expressed in the Twitter messages, the entity-based profiles provide a richer variety and allow for specifying users' interests on a more fine-grained level than category-based profiles as well as hashtag-based profiles.

We further investigate *LDA-based topic modeling* which infers users' interests based on latent topics. Therefore, we adopt Latent Dirichlet Allocation (LDA) introduced by Blei et al. [24]. LDA identifies latent topics in large document collections. With LDA, each document is typically represented as a a probability distribution over a set of topics, while each topic is a probability distribution over a set of terms. Since Twitter messages are limited to 140 characters and are thus very short, we aggregate all the tweets published by a user into a single document as also suggested in previous studies by Weng et al. [182] and Hong et al. [92]. The resulting documents would typically contain all the terms and text snippets that a user mentioned in her tweets. In this thesis, we will represent the documents via entities that are mentioned in the tweets in order to get rid of stopwords and supplemental chatter. For each document, LDA estimates a probability distribution over latent topics, each of which in turn refer to a probability distribution over entities. Given a user,

[1] http://www.opencalais.com

[2] http://d.opencalais.com/er/company/ralg-tr1r/23d07771-c50b-315b-8050-3cdaf47ac0d0.html

her interests are therefore modeled as either a mixture of latent topics or a mixture of entities. More formally, the user profile $P(u)$ for a given user $u \in U$, which is modeled via latent topics, is specified as follows:

$$P(u) = \{(z, p(z|d_u))|\forall z \in Z\} \tag{3.2}$$

where Z and U denote the set of latent topics generated given a set of documents and the set of users respectively. The number of latent topics has to be specified in advance and allows for adjusting the degree of specialization of the latent topics. d_u denotes a single document which aggregates all the entities extracted from microposts published by user u. And $p(z|d_u)$ is the probability of a latent topic z for a given document d_u.

Alternatively, the user profile $P(u)$, which is modeled as a mixture of entities, is specified as follows:

$$P(u) = \{(e, p(e|d_u))|\forall e \in E\} \tag{3.3}$$

where E denotes a set of distinct entities occurred in given documents. $p(e|d_u)$, which is computed according to the following formula, describes the occurrence probability of an entity e for document d_u.

$$p(e|d_u) = \sum_{z \in Z} p(e|z)p(z|d_u) \tag{3.4}$$

where $p(e_i|z)$ is the probability of entity e within topic $z \in Z$ and $p(z|d_u)$ is the probability of picking a topic z for the document d_u.

In Chapter 6, we will further investigate *LDA-based topic modeling* for constructing user profiles and evaluate the quality of LDA-based user profiles in the context of news recommender systems.

3.2.2 Enrichment

To overcome the shortness of microposts and further enrich the semantics of microblogging activities, we implemented several strategies that link microposts with external Web resources. We developed URL-based and content-based strategies that detect mappings between a micropost and possibly related external Web resources such as news articles. URL-based methods exploit hyperlinks mentioned in microposts to relate tweets with external resources. By exploiting the linkage between the micropost and the external Web resource, entities and categories extracted from that external resources are then propagated to the linked micropost. For example, two

category-based topic	micropost	external	entity-based topic	micropost	external
Sports	2	4	Francesca Schiavone	1	3
Technology_Internet	2	3	iPad	2	4
Entertainment_Culture	0	1	Fed Cup	1	2
			Italy	1	2
			Garageband	1	2
			Apple	1	1
			tennis	0	2
			French Open	0	2
			sportsman	0	1
			Hobart	0	1
			Jarmila Groth	0	1
			Flavia Penetta	0	1
			touch device	0	1

Table 3.4: Comparison of two design alternatives for semantic enrichment of microposts: (i) micropost-only-based, (ii) exploitation of external resources.

entities (*Fed Cup* and *Italy*) are extracted only from the textual content of micropost $t3$ from Table 3.2. By further exploiting a relevant news article that is linked via the shortened URL contained in the micropost, more entities (e.g., *tennis* and *Flavia Penetta*) can be obtained for constructing the user profiles. For those microposts that do not contain explicit links, the mappings are built by correlating timestamps and entities of Twitter messages with the ones of external Web resources.

In Table 3.4, we compare the topics of interests that are produced based on two design alternatives: micropost-only-based enrichment and exploitation of external Web resources via explicit URLs. For the latter, we propagate categories and entities extracted from the Web resources identified via the (shortened) URLs to the linked microposts. For the given example microposts in Table 3.2, the exploitation of external Web resources increases the variety of both category-based and entity-based topics. For example, one new category (*Entertainment_Culture*) and six new entities (e.g., *French Open*, *tennis*) are extracted from the related Web resources. In Chapter 4, we will explain in more detail the different strategies for linking microblogging activities with the external Web resources to enhance the semantics of microblogging and show that the best strategy achieves 80% accuracy. Furthermore, by conducting a large-scale analysis based on millions of microblogging activities, we will demonstrate that semantic enrichment provides a richer basis for microblogging-based user modeling and has a significant impact on the characteristics of user profiles.

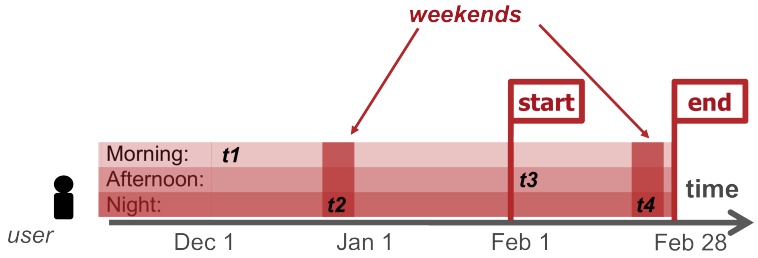


Figure 3.1: Temporal constraint for microblogging-based user modeling

3.2.3 Temporal Constraints

Figure 3.1 illustrates how the temporal constraints are taken into account when constructing user profiles based on microblogging activities. Firstly, the user profiles can be constructed based on the microblogging activities published during specific periods of time. Therefore our user modeling framework allows for the generation of long-term user profiles, which are based on the complete user history as well as short-term user profiles which are based on microblogging activities within a specific period of time. For example, given the microposts that are listed in Table 3.2, the messages $t3$ and $t4$ can be selected to construct a short-term user profile, which models a user's interests based on the microblogging activities published in the most recent month. In Chapter 6, we will further analyze the impact of such temporal constraint on the characteristics of user profiles. For example, does recent profile information approximates future profiles better than old profile information?

Secondly, our user modeling framework allows for investigating which temporal pattern occurs in the user profiles to better understand the temporal characteristics of user profiles. For example, weekend profiles are constructed based on the microposts $t2$ and $t4$, which are published during weekends (Saturday-Sunday). The exploitation of temporal patterns allows us to understand the temporal characteristics of user profiles and provide flexible options for constructing user profiles that meet the demands of various applications. In Chapter 5, we will investigate whether the users posting behavior during weekdays differs from the one during weekend for users from different cultural groups (e.g., Chinese and American users). We will also explore in Chapter 6 the differences between user profiles created on the weekends with those created during the week to detect temporal pattern that might help to improve personalization within certain time frames.

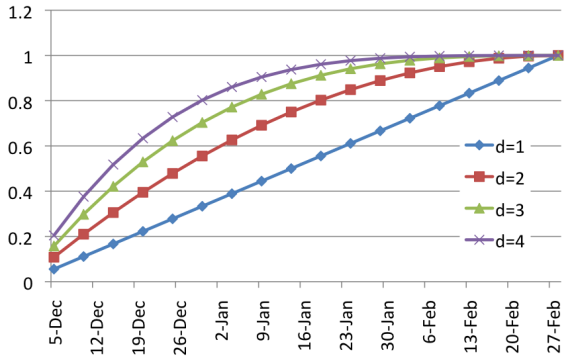


Figure 3.2: Impact of parameter d and timestamp $time_c$ on the weights of topic.

3.2.4 Weighting Schemes

Our user modeling framework allows for different weighting schemes. A straightforward approach is, for example, to count the number of occurrences of a concepts so that the weight $w(u,c,time_c)$ of a concept is determined by the number of microblogging activities in which the user u refers to topic c (see Equation 3.5).

$$w_{TF}(u,c) = \frac{|T_{tweets,u,c}|}{|T_{tweets,u}|} \tag{3.5}$$

where $T_{tweets,u,c}$ denotes the set of tweets published by user u that refer to topic c and $T_{tweets,u}$ denotes the set of tweets published by user u.

While the TF weighting scheme, which ignores the temporal information, provides a straightforward approach to measure the importance of users' interests, our user modeling framework further supports more advanced approaches which take time input into account. For example, Equation 3.6 describes a time-sensitive weighting function which dampens the occurrence frequency according to the temporal distance between the topic occurrence time and the given timestamp.

$$w(u,c,time_c) = \sum_{c \in T_{t,u,c}} (1 - \frac{|time - time_c|}{max_{time} - min_{time}})^d \tag{3.6}$$

In Equation 3.6, $T_{t,u,c}$ denotes the set of tweets that have been published by u and refer to the topic c. $time$ denotes the timestamp when the user profile is constructed. $time_c$ returns the timestamp of a given tweet where the given topic c is derived and max_{time} and min_{time} denote the highest (youngest) and lowest (oldest) timestamp of a micropost in $T_{tweets,u,c}$, for example: $max_{time} = max(\{time(t)|t \in T_{tweets,u,c}\})$. The parameter d is used to adjust the influence of the temporal distance. In Figure 3.2,

	$P_{m,TF}$	$P_{e,TF}$	$P_{e,TF,s}$	$P_{m,TS}$	$P_{e,TS}$
Francesca Schiavone	0.17	0.14	0.06	0.02	0.07
iPad	0.33	0.18	0.19	0.35	0.22
Fed Cup	0.17	0.09	0.13	0.21	0.13
Italy	0.17	0.09	0.13	0.21	0.13
Garageband	0.17	0.09	0.13	0.21	0.13
tennis	0	0.09	0	0.06	0.07
French Open	0	0.09	0.06	0	0.01
sportsman	0	0.05	0	0	0.01
Hobart	0	0.05	0	0.06	0.06
Jarmila Groth	0	0.05	0.06	0	0.06
Flavia Penetta	0	0.05	0.06	0	0.06
touch device	0	0.05	0	0.06	0.06

Table 3.5: Example entity-based user profiles constructed with different strategies.

we plotted the weights of a topic when varying the parameter d as well as the timestamp $time_c$ of the given micropost where the topic is derived. It shows that the closer the given topic occurs to the input $time$, the higher the corresponding weights will be, i.e. the time-sensitive function depicted in Equation 3.6 emphasizes the recently occurred topics. We further observe that the higher d is set, the higher the penalty of topics that occur with a high temporal distance to the input $time$ as the corresponding scores will be lower than for those topics for which $|time - time_c|$ is smaller.

Given the example microposts in Table 3.2, we constructed entity-based user profiles based on different combinations of design alternatives (see Table 3.5). In comparison to $P_{m,TF}$ that denotes the user profile constructed only based on the microblogging activities using the TF weighting schemes, the user profile $P_{e,TF}$, which is constructed by also exploiting the external Web resources, reduces the sparsity of the profile vector and varies the importance of some topics. For example, the topic Francesca Schiavone becomes more important in profile $P_{e,TF}$ than in profile $P_{m,TF}$ due to its more frequent occurrence in related news articles. Furthermore, the short-term user profile, denoted as $P_{e,TF,t}$, only consist of topics such as Garageband, which are derived from the microposts published within a certain time frame. We also observe that in the user profiles $P_{m,TF}$ and $P_{e,TS}$, both of which are constructed using time-sensitive weighting function (cf. Equation 3.6), the topics (e.g., Fed Cup, Italy, Garageband) derived from the later published microposts are weighted higher than in the profiles $P_{m,TF}$ and $P_{e,TF}$ that are constructed using TF weighting scheme. In Chapter 6 we will further analyze the impact of different design alternatives on the quality of constructed user profiles in the context of recommender systems. For example, in the context of a news recommender system, the time-sensitive weighting scheme may better characterize the recent interests of

a user than non-time-sensitive weighting schemes.

3.3 GeniUS - Generic User Modeling Library for the Social Semantic Web

Given the user modeling framework described above, we also implemented the framework and developed the so-called GeniUS library. Given textual user-generated content, GeniUS constructs semantic profiles that summarize the content of unstructured user data. With the contributions of GeniUS, we aim to (i) provide a flexible and extensible library that is able to serve different applications; (ii) produce semantically meaningful profiles to enhance the interoperability of profiles between applications; and (iii) customize the construction of user profiles according to the information needs of different applications. In this section, we first describe the architecture of GeniUS, including its four main modules, and then demonstrate how GeniUS can be applied to generate domain-specific user profiles.

3.3.1 Architecture of GeniUS

The architecture of GeniUS is presented in Figure 3.3. It is composed of four sequential modules, which process the given Social Web content to construct the profile. It consists of the *Item Fetcher*, *Enrichment*, *Topic Modeling & Weighting Function*, and *RDF Serialization*. Moreover, there are means for customizing the profile construction based on the demands of the application that is using GeniUS to obtain profiles: *Modeling Configuration* and *Filter*.

Item Fetcher Many Social Web services provide APIs that allow for collecting data to serve external applications. For example, Twitter exposes its data through the Twitter Streaming and Search API[3]. The main function of *Item Fetcher* is to collect raw content, either directly from the Social Web or from a local repository. Topic profiles can be generated if the item fetcher is configured – using keyword or SPARQL queries – to collect data that refers to a given topic while user profiles are generated if the fetched posts were published by te same user. We transform and represent the raw content based on a structured data model using the Semantically Interlinked Online Communication (SIOC) ontology [26]. Moreover, client applications can feed GeniUS with any types of SIOC items ranging from social bookmarking posts to (micro-)blog posts. The SIOC ontology provides a broad range of vocabulary con-

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

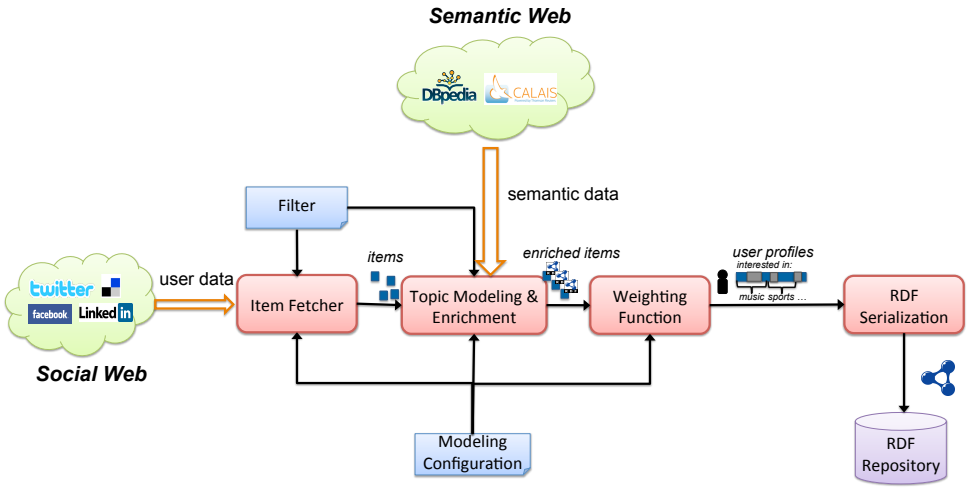


Figure 3.3: The architecture of GeniUS Library

cepts to describe information from the Social Web so that the *Item Fetcher* is highly flexible regarding the type of content it can handle. For example, to conduct our analysis and experiments on leveraging Twitter messages with GeniUS, we adopt *sioc:Post* to represent the tweets and *sioc:UserAccount* to describe the user account via which the messages were published.

Topic Modeling & Enrichment To better understand the semantics of the content collected via the *Item Fetcher*, we further extract relevant concepts from the textual content for modeling the topics of interests (see Section 3.2.1). This step is accomplished by using existing services. In particular, GeniUS provides adaptors for Zemanta[4] and SpotLight[5] [130]. Each extracted concept is identified with a unique, resolvable URI so that the meaning of a concept is well-defined and the semantically enriched Twitter posts as well as the generated profiles are well-connected to the Linked Open Data cloud[6]. In addition, we develop different strategies that link micropost with external Web resources to further enrich the semantics of microblogging activities (see Section 3.2.2). For example, we apply the enrichment component to connect microposts with external online news articles by exploring the explicit URLs contained in the microposts as well as comparing the semantics that are extracted in both microposts and news articles. For more detailed description and evaluation of different linkage enrichment strategies, we refer the reader to Chapter 4.

[4] http://developer.zemanta.com/
[5] http://dbpedia.org/spotlight
[6] http://www4.wiwiss.fu-berlin.de/lodcloud/state/

Weighting Function One of the main features of GeniUS is user profile construction for representing users' preferences and interests. Those types of profiles are essential for applications that aim for personalization. GeniUS mainly adopts the Vector Space Model to represent users' interests, i.e. a user profile is thus a set of weighted concepts. Therefore, we utilize the Weighted Interests Vocabulary and Weighting Ontology [33] as data model to represent user interest profiles .

The GeniUS library allows for different weighting functions to measure the importance and popularity of concepts in a profile ranging from straightforward strategies such as concept frequency (count the number of messages that refer to a concept) to more sophisticated strategies that compute a weight as a function of time, e.g., recently mentioned concepts are weighted stronger (see Section 3.2.4). Furthermore, client applications can specify their own weighting functions to customize the profile generation.

RDF Serialization The constructed user profiles can be outputted as RDF using FOAF [32] in combination with the Weighted Interest Vocabulary. Profiles can also be stored in an RDF repository. Client applications or end-users can thus retrieve RDF-based profiles from the repository and perform sophisticated RDF queries over the profiles. For example, GeniUS provides adapters for the Sesame RDF repository[7] to store the RDF profiles and allow for SPARQL queries.

In addition to the four modules mentioned above, which process the raw content and construct topic and user profiles, GeniUS also provides two configuration modules to enable a flexible modeling process that is required in order to generate domain- and application-specific profiles.

Modeling Configuration The aforementioned four GeniUS modules are exposed as interfaces (in Java) so that developers can easily extend GeniUS and implement new functions based on their needs. For example, we implemented time-sensitive weighting functions for applications that require more recent and dynamic characteristics of user profiles. The *Modeling Configuration* is used to configure which implementation for each module should be used in a GeniUS modeling process. With different combinations of module implementations, GeniUS has a variety of modeling alternatives to adapt to different applications.

Filter With the *Filter* feature, GeniUS is able to filter out irrelevant profile information and can construct user profiles that represent certain characteristics

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

of a user. We currently have implemented three types of filters: (i) filtering based on temporal constraints, (ii) keyword-based filtering, and (iii) semantic filtering. The fist strategy can filter out content collected via the *Item Fetcher* or can prevent the *Item Fetcher* from collecting items that do not fulfill the given temporal constraints. For example, one can restrict the Twitter message collection process to tweets that were published within a certain period of time (see Section 3.2.3). The second filtering module can filter out items that do not contain a given set of keywords or specifically collect items that match the given keywords. With the enrichment of the user-generated items with meaningful concepts, GeniUS can also perform semantic filtering to generate customized profiles that characterize a user in context of a specific domain. In the subsequent section, we will reveal how we use SPARQL queries as semantic filters on the semantically enriched items to filter out irrelevant noise before constructing the actual (weighted interest) profile.

We have released GeniUS[8] as a software library written in Java. With GeniUS one can easily specify the design alternatives for generating user profiles as well as implement news functions for each GeniUS module. Therefore, the user modeling process can be adapted to the demands of different applications. In Section 3.3.2, we will demonstrate how such adaptation can be applied to construct domain-specific user profiles.

3.3.2 Domain-specific User Profile Construction Using GeniUS

By utilizing filtering functionality, the library is able to build flexible user profiles for different application domains on demand. The constructed user profiles are represented in RDF with well-defined semantics. Given a short Twitter post like:

Awesome, love the Garageband for iPad http://bit.ly/fkj3YM #apple

we collect the content of this message and additional information such as the user identifier of the creator and the creation time via the Twitter Streaming API. Semantic Web vocabularies such as SIOC, Dublin Core and FOAF are applied to represent the Twitter message. We have also built a parser to process the content of the message and extract hashtags and URLs that are mentioned in a tweet. The extracted hashtags are identified using TagDef [9] as depicted in the following code snippet.

```
@prefix sioc: <http://rdfs.org/sioc/spec/> .
```

[8] http://wis.ewi.tudelft.nl/genius/

[9] http://www.tagdef.com

```
@prefix dcterms: <http://purl.org/dc/terms/> .
@prefix tagdef: <http://tagdef.com/> .

<http://twitter.com/bob/status/73748435752333312>
   a <sioc:Post> ;
   dcterms:created "2011-05-26T15:52:51+00:00" ;
   sioc:has_creator <http://twitter.com/bob> ;
   sioc:content "Awesome, love the Garageband for iPad http://bit.ly/fkj3YM #apple" ;
   sioc:links_to <http://bit.ly/fkj3YM> ;
   sioc:has_topic tagdef:apple .
```

The message is thus represented as *sioc:Post* and specifies metadata (e.g., *dcterms:created*, *sioc:has_creator*) as well as basic information about the content (e.g., *sioc:content*, *sioc:links_to*). *sioc:has_topic* is used to describe the semantic meaning of the content of the tweet. However, using TagDef does not allow for disambiguating the semantic meaning of the tweet and hashtag specifically. For example, *apple* may refer to the fruit or to the technology company. Hence, further semantic enrichment is required to specify the meaning of a Twitter message more accurately. Therefore, we perform named entity recognition and identify DBpedia concepts in tweets (using disambiguation functionality provided by DBpedia spotlight [130]). This allows us to further describe the topic of the tweet with further RDF statements using again the *sioc:has_topic* property:

```
@prefix sioc: <http://rdfs.org/sioc/spec/> .
@prefix dcterms: <http://purl.org/dc/terms/>
@prefix tagdef: <http://tagdef.com/> .
@prefix dbpedia: <http://dbpedia.org/resource/> .

<http://twitter.com/bob/status/73748435752333312>
   a <sioc:Post> ;
   dcterms:created "2011-05-26T15:52:51+00:00" ;
   sioc:has_creator <http://twitter.com/bob> ;
   sioc:content "Awesome, love the Garageband for iPad http://bit.ly/fkj3YM #apple" ;
   sioc:links_to <http://bit.ly/fkj3YM> ;
   sioc:has_topic tagdef:apple ;
   sioc:has_topic dbpedia:Apple_Inc. ;
   sioc:has_topic dbpedia:GarageBand ;
   sioc:has_topic dbpedia:IPad .
```

Given the semantic enrichment and the inferred additional RDF statements, we can disambiguate the meaning of the Twitter message: it refers to a software product (*dbpedia:GarageBand*) developed by the Apple company (*dbpedia:Apple_Inc.*) that is now available for the *iPad* device (*dbpedia:IPad*). By following the DBpedia URIs, applications can obtain further background information such as type information (e.g., *dbpedia:GarageBand* is of type *dbo:Software* and *yago:AudioEditors*[10]) or a list of persons that are involved in Apple (e.g., *dbo:keyPerson*).

[10]Here, *dbo* and *yago* refer to the DBpedia ontology (http://dbpedia.org/ontology/) and Yago ontology (http://dbpedia.org/class/yago/) respectively.

With the enriched concepts, GeniUS constructs profiles using a given weighting scheme. FOAF and the Weighted Interests Vocabulary are applied to describe the user and her preferences and interests into topics (based on the concepts that are referenced from the tweets). Since people publish Twitter messages on various different subjects, a generic approach, which considers all kinds of concepts that are referenced from the tweets, produces user profiles that contain a variety of topics. In the following example, the extract of the complete profile thus specifies topics of interests from different domains such as the music, software or movie domain:

```
@prefix foaf: <http://xmlns.com/foaf/0.1/> .
@prefix wi: <http://purl.org/ontology/wi/core#> .
@prefix wo: <http://purl.org/ontology/wo/core#> .
@prefix dbpedia: <http://dbpedia.org/resource/> .
@prefix genius: <http://persweb.org/genius#> .
<http://twitter.com/bob>
  a foaf:Person;
  wi:preference [
     a wi:WeightedInterest ;
     wi:topic dbpedia:Jazz ;
     wo:weight [
        a wo:Weight ;
        wo:weight_value 0.5889 ;
        wo:scale genius:Scale ]
     ] ;
  wi:preference [
     a wi:WeightedInterest ;
     wi:topic dbpedia:Second_Life ;
     wo:weight [
        a wo:Weight ;
        wo:weight_value 0.3114 ;
        wo:scale genius:Scale ]
     ] ;
  wi:preference [
     a wi:WeightedInterest ;
     wi:topic dbpedia:Short_film ;
     wo:weight [
        a wo:Weight ;
        wo:weight_value 0.3333 ;
        wo:scale genius:Scale ]
     ] ;
  wi:preference [
     a wi:WeightedInterest ;
     wi:topic dbpedia:GarageBand ;
     wo:weight [
        a wo:Weight ;
        wo:weight_value 0.1638 ;
        wo:scale genius:Scale ]
     ] ;  ...
```

The above profile depicts that the user is interested in jazz music (*dbpedia:Jazz*), short movies (*dbpedia:Short_film*) and software products (e.g., *dbpedia:Second_Life*). The higher the weight the higher the inferred interest in a concept. When applying the constructed profiles for a specific application domain, a drawback of such a

complete profile is that it lists also concepts that are possibly not relevant in the application context. For example, if a system aims to recommend software products to the above user then concepts such as *dbpedia:Jazz* or *dbpedia:Short_Film* might not add value to the profile while statements about the user's preference into software are more important. Utilizing the semantic filtering feature of GeniUS, application developers can specify a SPARQL query that describes what kind of topic-based profile a client application is seeking for:

```
SELECT DISTINCT ?t WHERE {
   ? <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <http://dbpedia.org/ontology/Software> }
```

Given such a SPARQL query, GeniUS will generate a customized profile where the concepts that do not belong to the software domain are filtered out (see below). The weight of the remaining concepts can be re-adjusted as well, for example, by normalizing the filtered profile.

```
<http://twitter.com/bob>
  a foaf:Person;
  wi:preference [
     a wi:WeightedInterest ;
     wi:topic dbpedia:Second_Life ;
     wo:weight [
        a wo:Weight ;
        wo:weight_value 0.4101 ;
        wo:scale genius:Scale
        ]
     ] ;
  wi:preference [
     a wi:WeightedInterest ;
     wi:topic dbpedia:GarageBand ;
     wo:weight [
        a wo:Weight ;
        wo:weight_value 0.2158 ;
        wo:scale genius:Scale
        ]
     ] ; ...
```

Using semantic filtering, applications can therefore utilize GeniUS to generate profiles that are well-adapted to their domain of interest. In Chapter 6, we will conduct further analysis and experiments to show the quality of such customized profiles in the context of recommender systems in different application domains.

3.4 Discussion

In this chapter, we investigated how to leverage microblogging activities for user modeling. We gave answers to the research questions raised at the begining of this

Design dimension	Hypothesis	Further validation
Topic modeling	*H1*: The entity-based topic modeling adds richer variety to user profiles than hashtag-based and category-based topic modeling.	Chapter 4, Chapter 6
Enrichment	*H2*: The exploitation of external Web resources creates more valuable user profiles than micropost-based strategies.	Chapter 4, Chapter 6
Temporal constraint	*H3*: The short-term profiles, which exploits the microblogging activities within certain time frames, differ from the long-term user profiles.	Chapter 6
Temporal constraint	*H4*: The temporal patterns (e.g., weekday vs weekend) impact the characteristics of user profiles.	Chapter 5, Chapter 6
Weighting scheme	*H5*: The time-sensitive weighting scheme characterizes the actual and concerns of a user better than the non-time-sensitive weighting scheme.	Chapter 6

Table 3.6: Hypotheses on the impact of different dimensions on the characteristics and quality of user profiles.

chapter by introducing TweetUM - a microblogging-based user modeling framework - that allows for inferring users' interests and generating semantically meaningful user profiles. TweetUM features a set of user modeling strategies that vary in four design dimensions: (i) *topic modeling* for inferring and representing the users' topics of interests, (ii) *enrichment* that exploits external Web resources to further enrich the semantics of user profiles, (iii) *temporal constraints* that are considered when constructing user profiles, and (iv) *weighting scheme* that measures the importance of topics in user profiles. By selecting and combining different design dimensions and design alternatives, our framework provides a variety of user modeling strategies that are able to serve different applications. We have already illustrated the influence of different design dimensions on the characteristics of constructed user profiles based on some example microposts (see Table 3.2) and produced hypotheses as shown in Table 3.6. Further analyses and experiments will be conducted to validate these hypotheses in the following chapters based on large amounts of microblogging data. For example, we analyze in Chapter 4 whether the semantic enrichment enhances the variety and quality of the microblogging-based user profiles. In Chapter 6, we investigate how the consideration of temporal constraints impact the personalized recommendation systems.

We developed GeniUS[11], which is an implementation of our Twitter-based user modeling framework. GeniUS implements set of strategies for the different design dimensions (e.g., enrichment, weighting schemes). Moreover, it features function-

[11] We make the GeniUS library publicly available via http://wis.ewi.tudelft.nl/genius/

ality for (i) collecting data from microblogging services, (ii) filtering microblogging data for different applications, and (iii) storing the user profiles. GeniUS allows for the generation of customized user profiles based on status messages people post on the microblogging platforms. Given a stream of messages, it generates topics and user profiles that summarize the stream according to domain- and application-specific needs which can be specified by the requesting party. Therefore, GeniUS can be applied in various application settings. In addition, GeniUS enriches user data with semantic information and constructs meaningful RDF-based profiles that are well-connected to the Linked Open Data cloud and therefore better support interoperability between applications that aim for personalization. In Chapter 6 we will further investigate the quality of profiles that are generated by different user modeling strategies for supporting various recommendations task ranging from product recommendations to more specific recommendations as required in a book or software product domain.

In the following chapters, we will introduce techniques for semantic enrichment of microblogging activities (see Chapter 4). We will further investigate how the user modeling framework can be applied into various applications including microblogging behavior analytics (see Chapter 5) and personalized recommender systems (see Chapter 6).

Chapter 4

Semantic Enrichment for Microblogging-based User Modeling

Having introduced the core user modeling framework, we now present and analyze techniques for semantic enrichment of microblogging activities, which provides the basis for the generation of semantically meaningful user profiles. The main contributions of this chapter have been published in [2].

4.1 Introduction

Learning and modeling the semantics of individual microblogging activities is important because the amount of microposts published each day is continuously growing so that users need support to benefit from microblogging information streams. For example, given the variety and recency of topics that people discuss on Twitter [58], user profiles that capture the semantics of individual tweets are becoming interesting for external applications on the Social Web. In order to enable Social Web applications to consume semantically meaningful representations of the users' microblogging activities, there is thus an urgent need to research user modeling strategies that allow for the construction of user profiles with rich semantics. Kwak et al. conducted a temporal analysis of trending topics in Twitter and show that hashtags are good indicators to represent events and trending topics [106]. Huang et al. analyze the semantics of hashtags in more detail and reveal that tagging in Twitter rather used to join public discussions than organizing content for future retrieval [93]. Laniada and Mika have defined metrics to characterize hashtags with

respect to four dimensions: frequency, specificity, consistency, and stability over time [109]. The combination of measures can help assessing hashtags as strong representative identifiers. Miles explored the retrieval of hashtags for recommendation purposes and introduced a method which considers user interests in a certain topic to find hashtags that are often applied to posts related to this topic [59]. Our research goes beyond hashtags as we extract and analyze semantics of microposts, which allows for a more comprehensive understanding of users' preferences.

To overcome the shortness of microposts, some research efforts have been made to explore external resources to further enrich the semantics of microblogging activities. Stankovic et al. map microposts to conference talks and exploit metadata of the corresponding research papers to enrich the semantics of single microposts [166]. Jadhav et al. present Twitris, which is a Semantic Web platform that connects event-related Twitter messages with other media such as Youtube videos and Google News [94]. In [106], the authors reveal that 85% of trending topics on Twitter are related to news. Mendes et al. developed Twarql that allows for capturing the semantics of major news events by detecting DBpedia entities from the content of microposts [129]. TwitterStand also analyzes the Twitter network to capture tweets that correspond to late breaking news [170]. Such analyses on certain news events, such as the election in Iran in 2009 [64] or the earthquake in Chile in 2010 [131], have also been conducted by other researchers. However, analyzing the feasibility of linking individual microblogging activities with news articles for enriching and contextualizing the semantics of user activities on Twitter to generate valuable user profiles for the Social Web – which is discussed in this chapter – has not been researched yet.

While the content of microposts can be objective in nature, it can also reflect emotions of individuals such as happiness or sadness. There has been a large amount of research that focus on detecting sentiment in user-generated online content like movie reviews [142]. Sentiment analysis over microblogging streams offers a number of possibilities to monitor a public's feelings as well as individual users' opinions about various topics such as trending news, products or events [40, 76, 176]. Go et al. introduce an approach for automatically classifying the sentiment of individual microposts as either positive or negative with respect to a query term [76]. In [55], authors analyze the microblogging streams to determine debate performance of the two candidates. Castellanos et al. present a system that allows for analyzing the evolution of sentiment in microblogging activities for a given topic or event [40]. In comparison to these applications that aim for monitoring and detecting the publics' feeling, we, in this chapter, investigate how to exploit emotions in the microblogging activities for modeling individual users' preferences and behavior. Furthermore, most research work has focused on analyzing only *positive* and *negative* sentiment. Such binary classification may miss important nuances in emotional states [51]. For

example, *Happy* and *Grateful* are both positive sentiments, but express two different emotional states. In this chapter we explore a set of more fine-grained emotional states to better grasp the human emotions expressed in the microblogging activities for individual users.

In this chapter, we will investigate how to model and learn the semantics of individual microblogging activities and answer the following research questions.

- How can we correlate microblogging activities with external Web resources in order to better understand the meaning of the microposts?
- How can we mine and exploit sentiment that is expressed in the microposts for semantic enrichment of microblogging activities?
- Which strategy allows for the highest accuracy for (i) correlating external Web resources with microposts and (ii) inferring the sentiment of microposts?
- How does the semantic enrichment impact the characteristics and quality of microblogging-based user profiles?

We will first in Section 4.2 introduce and evaluate the strategies for linking individual microblogging activities to external Web resources. We will further analyze the impact of exploiting linkage on the characteristics and quality of constructed user profiles. In Section 4.3 we will present an approach for identifying emotions from microposts and analyze the user profile construction based on the identified emotions.

4.2 Exploitation of Linkage for Microblogging-based User Modeling

Given the shortness of microposts, automatically inferring the semantic meaning of microblogging activities is a non-trivial problem. For example, posts such as "Interesting: http://bit.ly/iajV21 #politics" or "@nytimes this makes me scared" are even for humans difficult to understand without knowing the context. However, by following the links one can explore this context and grasp the semantics of the tweets. Many Twitter activities are related to news events. According to Kwak et al. more than 85% of trending topics in Twitter are related to news [106]. This observation motivates our idea of linking microposts with news articles to automatically capture and enrich the semantics of microblogging activities. Such relations between microposts and news further allow for capturing user interests regarding trending

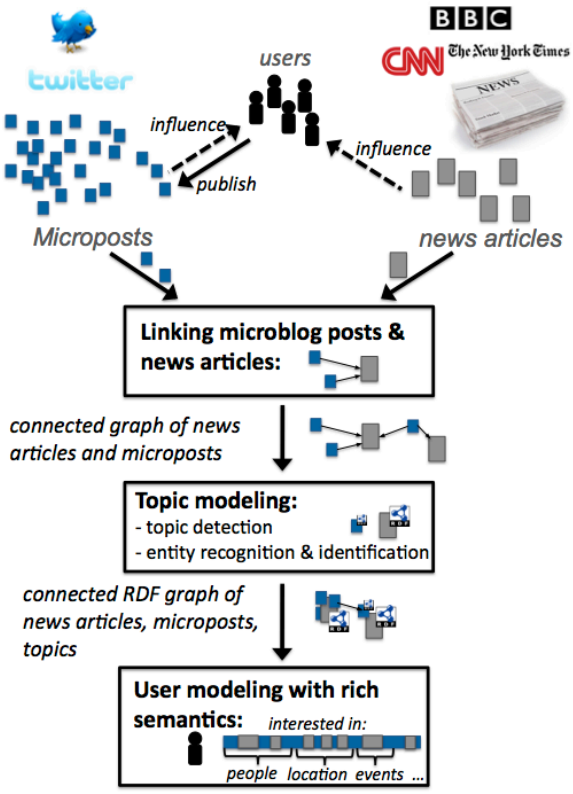


Figure 4.1: Generic architecture for linking tweets with news articles and constructing user profiles.

topics and support applications that require recent user interests like recommender systems for news or other fresh items.

Figure 4.1 visualizes the components of our approach for constructing user profiles with rich semantics based on Twitter posts. We relate Twitter messages with news articles and exploit the content of both tweets and news articles to derive the semantics of the users' microblogging activities. Therefore, we aggregate individual posts of Twitter users as well as news articles published by mainstream media such as CNN, BBC, or New York Times and propose the following components.

Linkage. The challenge of linking tweets and news articles is to identify those articles a certain Twitter message refers to. Sometimes, users explicitly link to the corresponding Web sites, but often there is no hyperlink within a Twitter message, which requires more advanced strategies. In Section 4.2.1 we introduce and evaluate different strategies that allow for the discovery of relations

between tweets and news articles.

Topic Modeling. Given the content of tweets and news articles, another challenge is to extract valuable semantics from the textual content. Further, when processing news article Web sites an additional challenge is to extract the main content of the news article. While RSS facilitates aggregation of news articles, the main content of a news article is often not embedded within the RSS feed, but is available via the corresponding HTML-formatted Web site. These Web sites contain supplemental content (*boilerplate*) such as navigation menus, advertisements or comments provided by readers of the article. To extract the main content of news articles we use BoilerPipe [101], a library that applies linguistic rules to separate main content from the boilerplate.

In order to support user modeling and personalization it is important to – given the raw content of tweets and news articles – distill topics and extract entities users are concerned with. As introduced in Section 3.2.1, we therefore utilize Web services provided by OpenCalais, which allow for the extraction of entities such as people, organizations or events and moreover assign unique URIs to known entities.

The connections between the semantically enriched news articles and Twitter posts enable us to construct a rich RDF graph that represents the microblogging activities in a semantically well-defined context.

User Modeling. Based on the RDF graph which connects Twitter posts, news articles, related entities and topics, we analyze user modeling strategies that are presented in Chapter 3, which create semantically rich user profiles describing different facets of the users (see Section 4.2.3).

Figure 4.2 further illustrates our generic solution by means of an example taken from our dataset: a user is posting a message about the election of the sportsman of the year and states that she supports Francesca Schiavone, an Italian tennis player. The Twitter message itself just mentions the given name *francesca* and indicates with a hashtag (*#sport*) that this post is related to sports. Hence, given just the text from this Twitter message it is not possible to automatically infer that the user is concerned with the tennis player. Given our linkage strategies, one can relate the Twitter message with a corresponding news article published by CNN, which details on the SI sportsman election and Francesca Schiavone in particular. Entity and topic recognition reveal that the article is about tennis (*topic:Tennis*) and Schiavone's (*person:Francesca_Schiavone*) success at French Open (*event:FrenchOpen*) and therewith enrich the semantics which can be extracted from the Twitter message itself (*topic:Sports*).

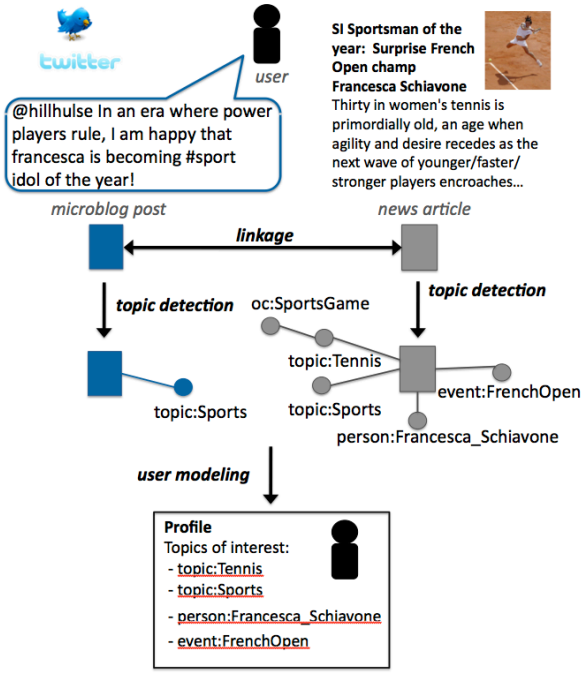


Figure 4.2: Example of linking tweets with news articles and constructing user profiles.

4.2.1 Linkage Discovery Strategies

The key idea of our approach to enrich the semantics of Twitter messages is based on relating individual tweets to news articles so that semantics extracted from news articles can be applied to clarify the meaning of tweets. In this section we introduce different strategies for linking Twitter posts with news articles that provide details on these posts. To evaluate the impact of these strategies on the semantic enrichment of Twitter posts we conduct an analysis on a large dataset gathered from Twitter and major news publishing Web sites.

The strategies, which we propose to find correlations between Twitter posts and external news resources, can be divided into URL-based strategies, which exploit interaction patterns and hyperlinks mentioned in tweets, and content-based strategies, which exploit the content of tweets and news articles. In the following definitions, T denotes the set of all tweets available in our dataset while N refers to the set of news articles that we collected from Web sites of mainstream news publishers.

URL-based Strategies

URLs (mostly short URLs shortened by services such as *bit.ly*) that are contained in tweets can be considered as indicators for news-related tweets. In particular, if a tweet contains a URL that points to an external news resource, there is a very high possibility that this tweet is closely related to the linked resource. Based on this principle we defined two URL-based strategies.

Definition 2 ***Strict URL-based strategy*** *If a Twitter post $t \in T$ contains at least one URL that is from certain mainstream news publishers and links to a news article $n \in N$, then we consider t and n as related: $(t,n) \in R_s$, where $R_s \subseteq T \times N$.*

For this strategy, we select BBC, CNN and the New York Times as the set of mainstream news publishers and apply URL-patterns to discover the corresponding tweets that point to news articles on these Web sites. N refers to the set of news articles from these threes mainstream news publishers. A potential drawback of the strict URL-based strategy is that it will miss relevant relations for Twitter messages that contains no URL. For example, if a user replies to a Twitter message that is according to the strict URL-based strategy related to a news article n then this reply message might be related to n as well. Based on this idea, we define a second URL-based strategy that is more flexible than the first one.

Definition 3 ***Lenient URL-based strategy*** *If a tweet $t_r \in T$ is a reply or re-tweet from another tweet $t \in T$, which contains at least one URL that is linked to a news article $n \in N$ authored by a mainstream news publisher, then we consider both t_r and t as being related to n: $(t_r,n) \in R_l, (t,n) \in R_l$, where $R_l \subseteq T \times N$.*

Hence, the lenient URL-based strategy extends the strict strategy with tweets that were published as part of an interaction with a tweet that is according to the strict strategy news-related so that $R_s \subseteq R_l$.

Content based Strategies

As tweets do not necessarily contain a URL, we propose another set of strategies that exploit the content of tweets and news articles to connect tweets with news. For example, the Twitter post about Francesca Schiavone in Figure 4.2 should be linked to the corresponding news article even though the tweet does not have a URL directly pointing to the article. We thus propose three further strategies that analyze the content of Twitter posts to allow for linkage between Twitter activities and news articles.

Definition 4 ***Bag-of-Words Strategy*** *Formally, a Twitter post $t_j \in T$ can be represented by a vector $\vec{t} = (\alpha_1, \alpha_2..\alpha_m)$ where α_i is the frequency of a word i in t and m denotes the total number of words in t. Each news article $n \in N$ is also represented as a vector $\vec{n} = (\beta_1, \beta_2..\beta_k)$ where β_i is the frequency of a word i in the title of the news article n and k denotes the total number of words in n.*

The bag-of-word strategy relates a tweet t with the news article n, for which the $TF \times IDF$ score is maximized: $(t,n) \in R_b$, where $R_b \subseteq T \times N$.

The bag-of-words strategy thus compares a tweet t with every news article in N and chooses the most similar ones to build a relation between t and the corresponding article n. $TF \times IDF$ is applied to measure the similarity. Given a Twitter post t and a news article n, the term frequency TF_i of a term i (with $\alpha_i > 0$ in the vector representation of t) is β_i, i.e. the number of occurrences of the word i in n. And IDF_i, the inverse document frequency is defined as follows.

$$IDF_i = 1 + log(\frac{|N|}{|\{n \in N : \beta_i > 0\}| + 1}) \tag{4.1}$$

where $|\{n \in N : \beta_i > 0\}|$ is the number of news articles, in which the term i appears. Given a tweet and a set of news articles, we rank the tweet-news pairs by calculating the similarity between between the tweet t and a news article n as follows.

$$sim(t,n) = \sum_{i=1}^{m} TF_i \cdot IDF_i \tag{4.2}$$

Given a ranking according to the above similarity measure, we select top ranked tweet-news pairs as candidates for constructing a valid relation. Following the realtime nature of Twitter, we also add a temporal constraint to filter out these candidates, for which the publishing date of the Twitter message and news article differs more than two days.

The bag-of-words strategy treats all words in a Twitter post as equally important. However, in Twitter, hashtags can be considered as special words that are important features to characterize a tweet [109]. For news articles, some keywords such as person names, locations, topics, etc. are also good descriptors to characterize a news article. Conveying these observations, we introduce hashtag-based and entity-based strategies for discovering relations between tweets and news articles. These strategies follow the idea of the bag-of-words strategy (see Definition 4) and differ in the way of representing news articles and tweets.

Definition 5 ***Hashtag-based strategy*** *The hashtag-based strategy represents a Twitter post $t \in T$ via its hashtags: $\vec{h} = (\alpha_1, \alpha_2..\alpha_m)$, where α_i is the number of occur-*

rences of a hashtag i in t and m denotes the total number of hashtags in t.

The hashtag-based strategy relates a tweet t (represented via its hashtags) with the news article n, for which the $TF \times IDF$ score is maximized: $(t,n) \in R_h$, where $R_h \subseteq T \times N$.

While the hashtag-based strategy thus varies the style of representing Twitter messages, the entity-based strategy introduces a new approach for representing news articles.

Definition 6 ***Entity-based strategy*** *Twitter posts $t \in T$ are represented by a vector $\vec{t} = (\alpha_1, \alpha_2..\alpha_m)$ where α_i is the frequency of a word i in t and m denotes the total number of words in t. Each news article $n \in N$ is represented by means of a vector $\vec{n} = (\beta_1, \beta_2..\beta_k)$, where β_i is the frequency of an entity within the news article, i is the label of the entity and k denotes the total number of distinct entities in the news article n.*

The entity-based strategy relates the Twitter post t (represented via bag-of-words) with the news article n (represented via the labels of entities mentioned in n), for which the $TF \times IDF$ score is maximized: $(t,n) \in R_e$, where $R_e \subseteq T \times N$.

Entities are extracted by exploiting OpenCalais (see Section 3.3). For the hashtag- and entity-based strategies, we thus use Equation 4.2 to generate a set of candidates of related tweet-news pairs and then filter out these pairs, which do not fulfill the temporal constraint that prescribes that the tweet and news article should be published within a time span of two days. Such temporal constraints may reduce the recall but have a positive effect on the precision as we will see in our analysis below.

4.2.2 Evaluation of Linkage Discovery

To analyze the impact of the strategies on semantic enrichment of Twitter posts, we evaluate the performance of the strategies with respect to coverage and precision based on a large data corpus which we crawled from Twitter and three major news media sites: BBC, CNN and New York Times.

Data Collection and Characteristics

Over a period of three weeks we crawled Twitter information streams via the Twitter streaming API. We started from a seed set of 56 Twitter accounts (U_n), which are maintained by people associated with one of the three mainstream news publishers, and gradually extended this so that we finally observed the Twitter activities of

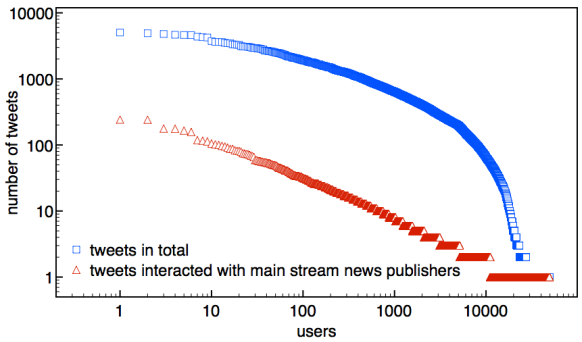


Figure 4.3: Number of tweets per user $u \in U_u$ as well as the number of interactions (re-tweeting or reply activities) with Twitter accounts maintained by mainstream news media.

48,927 extra users (U_u), who are not explicitly associated with BBC, CNN or the New York Times. The extension was done in a snowball manner: we added users to U_u, who interacted with another user $u \in U_n \cup U_u$. The 56 Twitter accounts closely related to mainstream news media enable publishers of news articles to discuss their articles and news events with the Twitter audience. For the 48,927 *casual users* we crawled all Twitter activities independently whether the activity was part of an interaction with the Twitter accounts of the mainstream news media or not. In total we thereby obtained more than 3.3 million tweets.

Figure 4.3 shows the number of tweets per user and depicts how often these users interacted with a Twitter account associated with mainstream media. The distribution of the number of tweets per user shows a power-law-like distribution. For many users we recorded less than 10 Twitter activities within the three week observation period. Less than 500 users were highly active and published more than 1000 tweets. Further, the majority (more than 75%) of users interacted only once with a news-related user $u \in U_n$. We observed only nine users, who re-tweeted or replied to messages from news-related users more than 100 times and identified one of these users as spam user, who just joined the discussion to promote Web sites.

To connect tweets with news articles, we further crawled traditional news media. Each of the three mainstream news publishers (BBC, CNN, and New York Times) also provides a variety of news channels via their Web site. These news channels correspond to different news categories such as politics, sports, or culture and are made available via RSS feed. We constantly monitored 63 different RSS feeds from the corresponding news publishers and crawled the main content as well as supplemental metadata (title, author, publishing data, etc.) of more than 44,000 news articles.

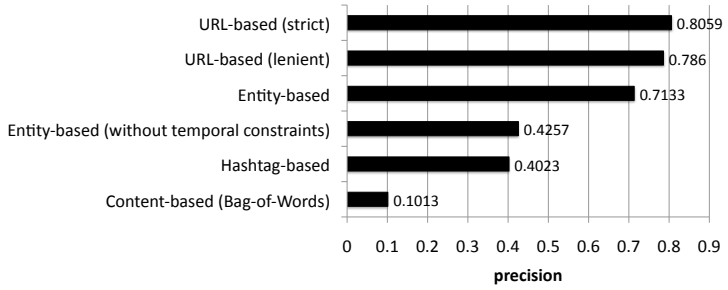


Figure 4.4: Precision of different strategies for relating tweets with news articles.

Experimental Results

In order to evaluate the accuracy of the different strategies for relating tweets with news articles, we randomly selected tweet-news pairs that were correlated by a given strategy and judged the relatedness of the news article and the tweet message on a scale between 1 ("not related") and 4 ("perfect match"), where 2 means "not closely related" and 3 denotes "related" tweet-news pairs. For example, given a Twitter message about Francesca Schiavone's victory at the French Open 2010, we considered news articles that report about this victory as a "perfect match" while news articles about Francesca Schiavone, for which this victory is not the main topic but just mentioned as background information were considered as "related". In total, we judged 1427 tweet-news pairs where each of the strategies depicted in Figure 4.4 was judged at least 200 times. For 85 pairs (5.96%) we were not able to decide whether the corresponding Twitter posts and the news article are related. We considered these pairs as "not related" and tweet-news relations, which were rated at least with 3 as truly related.

Given this ground truth of correct tweet-news relations, we compare the precision of the different strategies, i.e. the fraction of correctly generated tweet-news relations. Figure 4.4 plots the results and shows that the URL-based strategies perform best with a precision of 80.59% (strict) and 78.8% (lenient) respectively. Notice that the precison of URL-based strategies is not 100%. We observe that for some tweets-news pairs discovered via URL-based strategies, although the tweets contain (shortened) URLs pointing to news articles, the content of news articles is not related to the Twitter messages themselves. The naive content-based strategy, which utilizes the entire Twitter message (excluding stop-words) as search query and applies TFxIDF to rank the news articles, performs worst and is clearly outperformed by all other strategies. It is interesting to see that the entity-based strategy, which considers the publishing date of the Twitter message and news article, is nearly as good as the lenient URL-based strategy and clearly outperforms the hashtag-based

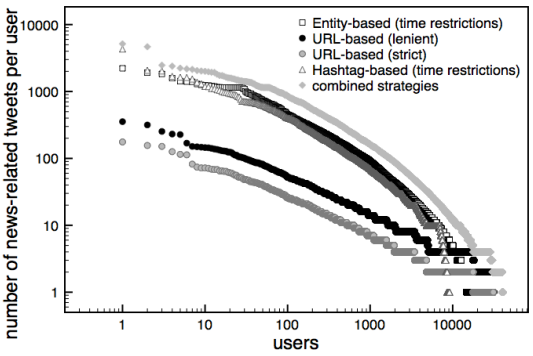


Figure 4.5: Number of tweets per user according to the different strategies related to news articles.

strategy, which uses the temporal constraints as well. Even without considering temporal constraints, the entity-based strategy results in higher accuracy than the hashtag-based strategy. We conclude that the constellation/set of entities mentioned in a news article and Twitter message correspondingly, i.e. the number of shared entities, is a good indicator of relating tweets and news articles.

Figure 4.5 shows the coverage of the strategies, i.e. the number of tweets per user, for which the corresponding strategy found an appropriate news article. The URL-based strategies, which achieve the highest accuracy, are very restrictive: for less than 1000 users the number of tweets that are connected to news articles is higher than 10. The coverage of the lenient URL-based strategy is clearly higher than for the strict one, which can be explained by the number of interactions with Twitter accounts from mainstream news media (see Figure 4.3). The hashtag-based and entity-based strategies even allow for a far more higher number of tweet-news pairs. However, the hashtag-based strategy fails to relate tweets for more than 79% of the users, because most of these people do not make use of hashtags. By contrast, the entity-based strategy is applicable for the great majority of people and, given that it showed an accuracy of more than 70% can be considered as the most successful strategy.

Combining all strategies results in the highest coverage: for more than 20% of the users, the number of tweet-news relations is higher than 10. In the next section we will show that given these tweet-news relations we can create rich profiles that go beyond the variety of profiles, which are just constructed based on the tweets of the users.

4.2.3 Analyzing User Profile Construction based on Linkage Discovery

Based on the linkage of Twitter activities with news articles, we can exploit the semantics embodied in the news articles to create and enrich user profiles. In this section, we first present approaches for user modeling based on Twitter activities and then analyze the impact of exploiting related news articles for user profile construction in Twitter.

User Modeling Strategies

In Chapter 3, we proposed several possibilities to model the topics of interests for individual users. In this study we focus on two types of profiles: entity-based and category-based profiles. An entity-based profile models a user's interests for a given set of entities such as persons, organizations, or events and can be defined as follows.

Definition 7 (Entity-based profile) *The* entity-based profile *of a user $u \in U$ is a set of weighted entities where the weight of an entity $e \in E$ is computed by a certain strategy w with respect to the given user u.*

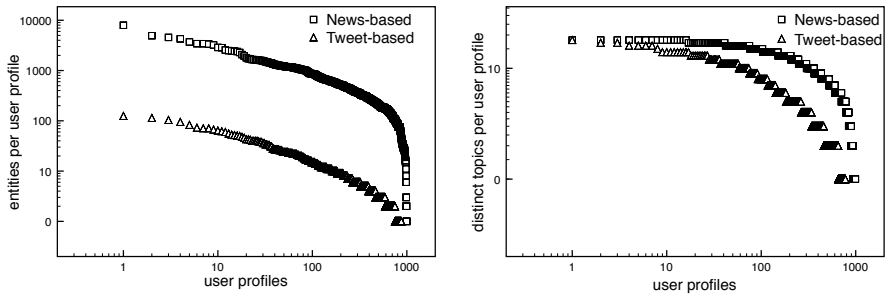
$$P(u) = \{(e, w(u,e)) | \forall e \in E\}$$

$w(u,e)$ is the weight that is associated with an entity e for a given user u. E and U denote the set of entities and users respectively.

In Twitter, a naive strategy for computing a weight $w(u,e)$ is to count the number of u's tweets that refer to the given entity e. $|P(u)|$ depicts the number of distinct entities that appear in a profile $P(u)$. While entity-based profiles represent a user in a detailed and fine-grained fashion, category-based profiles describe a user's interests for categories such as sports, politics or technology that can be specified analogously (see Definition 8).

Definition 8 (Category-based profile) *The* category-based profile $P_T(u)$ *of a user $u \in U$ is a set of weighted categories where the weight of a category is computed by a certain strategy with respect to the given user u.*

From a technical point of view, both types of profiles specify the interest of a user into a certain URI, which represents an entity or category respectively. Given the URI-based representation, the entity- and category-based profiles become part of the Web of Linked Data and can therewith not only be applied for personalization purposes in Twitter (e.g., recommendations of tweet messages or information streams to follow) but in in other systems as well. For the construction of entity- and category-based profiles we consider and compare the following two strategies.



(a) Entity-based profiles (b) category-based profiles

Figure 4.6: Comparison between tweet-based and news-based user modeling strategies for creating (a) entity-based profiles and (b) category-based profiles.

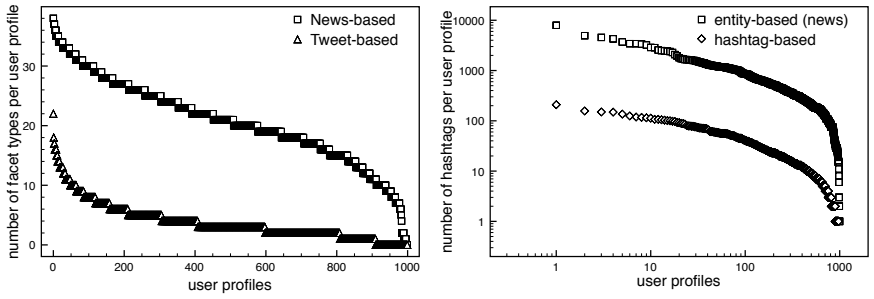
Tweet-based The tweet-based baseline strategy constructs entity- and category-based user profiles by considering only the Twitter messages posted by a user, i.e. the first step of our user modeling approach depicted in Figure 4.1 is omitted so that tweets are not linked to news articles. Entities and categories are directly extracted from tweets using OpenCalais. The weight of an entity corresponds to the number of tweets, from which an entity was successfully extracted, and the weight of a category corresponds to the number of tweets, which were categorized with the given category.

News-based The news-based user modeling strategy applies the full pipeline of our architecture for constructing the user profiles (see Figure 4.1). Twitter messages are linked to news articles by combining the URL-based and entity-based (with temporal restrictions) strategies introduced in Section 4.2.1 and entities and categories are extracted from the news articles, which have been linked with the Twitter activities of the given user. The weights correspond again to the number of Twitter activities which relate to an entity and category respectively.

Our hypothesis is that the news-based user modeling strategy, which benefits from the linkage of Twitter messages with news articles, creates more valuable profiles than the tweet-based strategy.

Analysis and Evaluation

To validate our hypothesis we randomly selected 1000 users (from U_u) and applied both strategies to create semantic user profiles from their Twitter activities. Fig-



(a) User profile facets (b) Hashtag- vs. entity-based profiles

Figure 4.7: Comparison between tweet-based and news-based user modeling strategies with respect to (a) the variety of facet types available in the user profiles (example facet types: person, event, location, product) and (b) number of distinct hash tags and entities per profile.

ure 4.6 compares the number of distinct entities and categories available in the corresponding profiles ($|P(u)|$). Even though the number of Twitter activities, which can be linked to news articles, is smaller than the total number of Twitter activities of a user (cf. Fig. 4.3, Fig. 4.5), the number of entities and categories available in the profiles generated via the news-based strategy is higher than for the tweet-based approach. Regarding the entity-based profiles this difference is higher than for the category-based profiles, because each Twitter message and news article is usually categorized with one category at most whereas for the number of entities there is no such limit. News articles provide much more background information (a higher number of entities) than Twitter messages and thus allow for the construction of more detailed entity-based user profiles.

Further, the variety of the entity-based profiles generated via the news-based strategy is much higher than for the tweet-based strategy as depicted in Figure 4.7(a). For the tweet-based strategy, more than 50% of the profiles contain just less than four types of entities (mostly persons and organizations) while for the news-based strategy more than 50% of the profiles reveal interests in more than 20 types of entities. For example, they show that users are – in addition to persons or organizations – also concerned with certain events or products. The news-based strategy, i.e. the complete user construction pipeline proposed in Figure 4.2, thus allows for the construction of profiles that cover different facets of interests which increases the number of applications that can be built on top of our user modeling approaches (e.g., product recommendations).

Related research stresses the role of hashtags for being valuable descriptors [59, 93, 109]. However, a comparison between hashtag-based profiles and entity-based

profiles created via the news-based strategy shows that for user modeling on Twitter, hashtags seem to be a less valuable source of information. Figure 4.7(b) reveals that the number of distinct hashtags available in the corresponding user profiles is much smaller than the number of distinct entities that are discovered with our strategy, which relates Twitter messages with news articles. Given that each named entity as well as each category of an entity- and category-based user profile has a URI, the semantic expressiveness of profiles generated with the news-based user modeling strategy is much higher than for the hashtag-based profiles. In Chapter 6, we will further investigate the impact of different user modeling strategies on the characteristics of constructed user profiles and the performance of personalized recommendations.

4.3 Exploitation of Emotion for Microblogging-based User Modeling

In the previous section, we analyzed methods for linking microposts with external Web resources so that content from those external resources can be used to better understand the (semantic) meaning of microposts. Semantically meaningful concepts are extracted via existing tools like OpenCalais to model the topics of interests for individual users. On the one hand, such concepts, which are mostly objective, reveal the facts (e.g., persons, locations and events) in the content of microposts. On the other hand, microblogging services like Twitter are increasingly used by people to share personal emotions and opinions on topics of interests. For example, given the following microposts published by two individual users:

Awesome, love the new released iPad #apple #excited.

No video camera on iPad :-(http://bit.ly/cnlBJZ #disappointed.

By extracting relevant concepts from the textual content, we may infer that these two users share a similar topic of interests (*iPad*). However, further examining the emotions in the microposts reveals that the two users apparently expressed different opinions on the same topic. While the positive opinion expressed in the first micropost indicates that the user has a keen interest in *iPad*, the second piece of text expresses clearly the user' negative opinion on the same product. Understanding emotions is beneficial for better modeling users' preferences and behavior in the microblogging sphere. However, the shortness of microposts and informal language used in the text make it challenging to identify emotions in the microblogging ac-

tivities. In this section we investigate strategies that exploit emotions expressed in the microposts for further enrichment of microblogging activities. Therefore, we

- propose an approach which utilizes conventional markers such as hashtags and emoticons to automatically label emotions in large Twitter datasets,
- explore and evaluate a variety of features to identify human emotions in microblogging streams, and
- investigate the characteristics of emotion-based user profiles that are inferred based on microblogging activities.

4.3.1 Emotions in Microposts

The classification of human emotions has been addressed by the research in psychology. For example, Ekman propose six-basic emotion classes including *happy*, *sad*, *anger*, *fear*, *surprise*, and *disgust*, as being common across cultures [60]. Earlier research efforts explored methods to identify positive and negative sentiments expressed in social media data such as product reviews [143] or blog posts [77]. However, such binary classification may miss the rich diversity of human emotions which are revealed in microblogging activities. Some research initiatives exploit more fined-grained classification of emotions on microblogging data. Purver et al. conducted experiments to classify Twitter messages into the six-basic emotions as mentioned above [150]. In [180], Wang et al. propose one more basic emotion, *grateful*, which is not covered by Ekman's classification. By manually checking a set of randomly picked tweets, we further added three more emotion classes, which are *grateful*, *like* and *dislike*, to Ekman's basic emotion classes. Therefore, we use an extended set of Ekman's basic emotions. which includes 9 types of emotions in total, and thus interpret the task of emotion identification as multi-class classification problem.

In order to generate groundtruth for evaluating the performance of emotion classification experiments, we created a collection of Twitter messages which are labeled with a set of emotion classes. Given the dataset described in Section 4.2, we selected a sample of 1619 active Twitter users, who have published at least 20 posts in the previous dataset, and further crawled their Twitter activities for another 2 months. In total we thereby obtained more than 4 million microblogging activities. For such a large amount of Twitter data, it's impossible to manually label all the messages. Therefore, we propose an approach that utilize conventional markers including hashtags, emoticons and Internet slang words to automatically label all Twitter messages in our dataset with a set of predefined emotion classes. Table 4.1

emotion class	conventional markers
Happy	:-) :) ;-) :D :d :P 8) lol haha #happy #happiness
Sad	:-(:(;-(:-¡ :(#sad #sadness
Anger	:-@ :@ #angry #anger
Fear	:— :-o :-O #scared #fear
Surprise	:s :S omg wow #surprised #surprise
Disgust	:$ +o(#disgusted #disgust
Like	like #love #enjoy
Dislike	don't like terrible
Grateful	thankful thank

Table 4.1: Conventional markers used to label emotion in microposts

lists the conventional markers used for automatically labeling tweets. We created a list of emoticons and hashtags based on a previous study conducted by Purver et al. [150], where the authors concluded that such conventional markers are reliable in labeling emotions in Twitter messages. De Choudhury et al. conducted a user study via Amazon Mechanical Turk[1], which shows that using affective words as hashtags (e.g, #happy) captures users' emotions with an accuracy of 83% [51]. Additionally, By examining the aforementioned set of randomly selected tweets, we further observed that some popular Internet slang words (e.g., 'lol', 'omg') can also be considered as labels of some emotion classes (see Table 4.1).

By utilizing a set of conventional markers listed in Table 4.1, we processed our Twitter data to automatically label the Twitter messages with predefined nine emotion classes. Several filtering heuristics were developed to help with the labeling process. We removed all URLs and user names in the Twitter messages before we started labeling these tweets. We also discarded tweets that contain less than five words, since they may not provide sufficient context to infer emotions as well as users' preferences. In the end, we collected 248,491 labeled tweets, which were published by 1,537 users, from more than 4 million microposts.

4.3.2 Emotion Classification Strategies

Given the labeled Twitter dataset, we conducted experiments on identifying emotions in Twitter messages by exploiting user-level features. Our assumption is that individual users' topics of interests are associated with different emotion classes and can thus be used to identify emotions expressed in Twitter messages. Therefore, we propose a user modeling approach that allows for generating emotion-based user profiles that reflect individual users' interests which are associated with various

[1] https://www.mturk.com

emotion facets. The generic model of emotion-based user profiles is specified in Definition 9.

Definition 9 ***Emotion-based User Profile*** *The emotion-based profile $P(u,e)$ of a user $u \in U$ represents the user's interests for a specific emotion $e \in E$. $P(u,e)$ is a set of weighted concepts where with respect to the given user u for a topic $c_e \in C_e$ its weight $w(u,c)$ is computed by a certain weighting function w.*

$$P(u,e) = \{(c_e, w(u,c_e))|\forall c_e \in C_e\} \tag{4.3}$$

Here, U denote the set of users. C_e denotes the set of topics derived from the Twitter messages that are labeled as emotion class e. With $\vec{p}(u,e)$ we refer to $P(u,e)$ in its vector space model representation and the values of all elements in that vector are normalized to make the sum of values to 1.

According to Definition 9, one has to decide what kind of concepts $c \in C_e$ are used to model topics of interests. In Chapter 3, we already investigated different topic modeling strategies for inferring users' preferences from microblogging streams. In this section we employ three of them to construct the emotion-based user profiles: hashtag-, category-, and entity-based topic modeling strategies. While the category-based and entity-based strategies allow for the generation of semantically meaningful user profiles, they may suffer from sparsity problem for constructing emotion-based user profiles due to the amount of Twitter messages that express emotions. In Section 4.2, we use OpenCalais which is capable of extracting named entities such as people and location from microposts. We also showed that the named entity extraction is beneficial for linking microposts to relevant external Web resources. In this section, we try to analyze more generic topics that users expressed in emotional microposts. Therefore, we further developed two more strategies for modeling topics of a user's interests: *WordNet-based topic modeling* and *synset-based topic modeling*. Both strategies utilize WordNet [137], which is a large lexical database of English, for constructing user profiles based on WordNet concepts extracted from Twitter messages. For the *WordNet-based topic modeling* strategy, we first clean the text of Twitter messages by removing URLs and common English stop words and then tonkenize each Twitter message into single words as queries for searching the WordNet database. Only the nouns and adjectives are used to model the topics that a user is interested in. WordNet also grounds words into set of synonyms called synsets. Therefore, each synset consists of a list of semantically related words. Given the topics represented using the nouns and adjectives that are found via the *WordNet-based topic modeling strategy*, the *synset-based topic modeling* strategy further uses the synonyms in their synsets to model the topics of interests for individual users.

	hashtag	category	entity	Wordnet	synset	mixed	all
Correctly Classified Instances	43.17%	43.03%	44.53%	64.19%	58.33%	62.04%	70.24%
Kappa statistic	0.0030	0	0.031	0.4332	0.3279	0.3927	0.5337
Mean absolute error	0.1492	0.1477	0.1485	0.1224	0.1371	0.124	0.0981
Root mean squared error	0.2731	0.2717	0.2725	0.2443	0.2609	0.2467	0.2179

Table 4.2: Summary for classification of emotions.

Based on the dataset described in Section 4.3.1, we construct different types of user profiles that differ with respect to the strategy for modeling topics of interests: hashtag-, category-, entity-, WordNet- and synset-based. Our main goal is to investigate the impact of different topic modeling strategies on the performance of emotion classification experiments. We thus use logistic regression [84] as the classifier in our experiments and compare the performance of classification by using various feature sets. For training the classifier, we extract features based on the similarity between Twitter messages and emotion-based user profiles. More specifically, given a tweet t published by a user u, we create the emotion-based user profile $P(u,e)$ of user u for a emotion class e using certain topic modeling strategy. Given a specific topic modeling strategy, we then generate for each user in our dataset nine emotion-based user profiles for the predefined emotion classes respectively (see Table 4.1). The tweet profile $P(t)$ is constructed with the same topic modeling strategy. For each pair of tweet profile $P(t)$ and a emotion-based profile $P(u,e)$, we calculate the cosine similarity between $P(t)$ and $P(u,e)$ as follows:

$$sim_{cosine}(\vec{p}(u,e),\vec{p}(t)) = \frac{\vec{p}(u,e)\cdot\vec{p}(t)}{||\vec{p}(u,e)||\cdot||\vec{p}(t)||} \tag{4.4}$$

Here $p(t)$ denotes the vector space representation of tweet profile $P(t)$, which is constructed via the same set of concepts as in user profile $P(u,e)$. To construct the user profiles as well as the tweet profiles, we apply the term-frequency weighting scheme, where the weight of a topic in a profile is determined by the number of occurrences of this topic (cf. Section 3.2.4). Based on different topic modeling strategies, we built a set of feature extractors that generate different types of features including: (i) hashtag-based, (ii) category-based, (iii) entity-based, (iv) wordnet-based and (v) synset-based. Throughout, all the labels were removed before feature extraction in the classification experiments, i.e. labels were not used as features.

4.3.3 Evaluation of Emotion Classification

We conducted the classification experiments using Weka [84]. The performance of emotion classification was evaluated using 10-fold cross-validation and is summa-

Class	TP Rate	FP Rate	Precision	Recall	F_1	ROC Area
Happy	0.800	0.371	0.546	0.800	0.649	0.744
Sad	0.194	0.003	0.489	0.194	0.278	0.890
Anger	0.141	0.003	0.447	0.141	0.215	0.887
Fear	0.256	0.003	0.621	0.256	0.363	0.891
Surprise	0.070	0.005	0.399	0.070	0.120	0.838
Disgust	0.015	0.000	1.000	0.015	0.029	0.726
Like	0.735	0.163	0.773	0.735	0.753	0.860
Dislike	0.244	0.005	0.533	0.244	0.334	0.904
Grateful	0.212	0.011	0.686	0.212	0.324	0.827
Weighted Avg.	0.642	0.205	0.651	0.642	0.615	0.817

Table 4.3: Detailed results for classification of emotion with WordNet-based feature.

rized in Table 4.2. We explore five different types of profiles (*hashtag-*, *category-*, *entity-*, *WordNet-* and *synset-based*) to extract features from Twitter messages. Using *WordNet-based* feature for the training, the classifier achieves an accuracy that equals to 64.19%, which is clearly higher than the classifier trained using features based on hashtag-based, category-based or entity-based profiles. In comparison to the *WordNet-based* feature, the *synset-based* feature, which utilize the WordNet synsets to further enrich the semantics of microposts, does not increase the accuracy of the classifier. The results show that the classifier using *synset-based* feature achieves 58.33% accuracy, which is lower than the *WordNet-based* feature, but still significantly outperforms the *hashtag-*, *category-*, and *entity-based* features. Using the *hashtag-based* feature, which is extracted based on the syntactic characteristics of Twitter messages, the classifier only achieves 43.17% accuracy. The accuracy of the classifier using the *category-based* feature achieves the lowest accuracy. By extracting more semantics from tweets to construct the user profiles and tweet profiles, the *entity-based* feature slightly improves the performance of classifier, achieving an accuracy that equals to 44.53%. Combining all features, the classifier achieves the highest accuracy (70.2%). We further evaluated the classifier using the mixed profiles where all the five types of topics, ranging from hashtags to Wordnet synsets, are utilized to construct the user profiles. The results reveal that the features based on such mixed profiles does not increase the performance of emotion classification. Instead, it is more valuable to construct the features based on separate topic modeling strategies.

In Table 4.3, we further show the detailed accuracy by class from the classifier using WordNet-based feature. The performance of classification varies very much in types of emotion. The *Like* class has the best performance with respect to F_1 measure. We observed that the classifier obtains very good results for the prediction of *Like* instances, achieving the highest *Precision* score and the second highest

type of emotion	proportion of posts
Happy	42.16%
Sad	4.63%
Anger	4.14%
Fear	5.47%
Surprise	7.32%
Disgust	0.22%
Like	12.17%
Dislike	6.32%
Grateful	17.57%

Table 4.4: Emotion expressed in overall posts.

Recall score across the classes. An F_1 measure equivalent to 0.753 illustrates that specially for the *Like* class the classifier obtains a good balance for the precision-recall tradeoff. The *Happy* class, which is also a positive emotion, is ranked second best class with respect to $F1$ (0.649). Furthermore, the F_1 score of the *Disgust* class is only 0.029. Note that the *Disgust* class has the least number of instances in our dataset. Both the user profiles and tweet profiles for this class tend to be very sparse, which might make it difficult to extract useful features since all the features are extracted by calculating the similarity between user profiles and tweet profiles. Additionally, the *Recall* of the negative emotion classes such as *Fear*, *Dislike* is lower than the positive emotion classes such as *Happy* and *Like*. For example, while the *Fear* class has higher *Precision* (0.621) than the *Like* class (0.546), the *Recall* of class *Fear* is 0.256, much lower than the *Recall* of class *Like* (0.8).

4.3.4 Analyzing Emotion-based User Profiles

In Section 4.3.3, we proposed an approach to constructing microblogging-based user profiles that allows us to model individual users' interests for different types of emotions. Based on the Twitter dataset where the Twitter messages have been labeled with a set of predefined emotion classes (see Section 4.3.1), we conduct an analysis to investigate the influence of emotions on the characteristics of constructed emotion-based user profiles.

The first research question investigated in our analysis is to what extent individual users express different types of emotions in their Twitter messages. Table 4.4 overviews the proportion of Twitter messages that have been labeled as one of nine

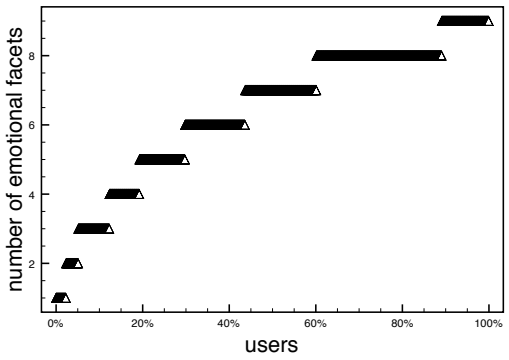


Figure 4.8: Number of distinct emotion facets for individual users

predefined emotion classes. The portion of tweets (42.16%) that have been labeled as *happy* is clearly higher than other emotion classes. Only 0.22% of the tweets express emotion *disgust*. We consider three types of emotions (*happy*, *like* and *grateful*) as positive emotions. Table 4.4 shows that 72% of the tweets have been labeled as positive emotions. Therefore, users tend to post more positive messages than negative ones on Twitter. In Chapter 5, we will conduct sentiment analysis on a larger dataset collected from different microblogging platforms to further study the positive and negative emotions in microposts. To analyze the different types of emotions revealed in the microposts for individual users, we plotted in Figure 4.8 for each user in our user sample (1,537 users), the number of distinct emotion classes labeled in the microposts. More than 80% of users expressed in their Twitter messages at least four different types of emotions out of nine. In addition, about 11% of users expressed all the nine types of emotions in their tweets. The results indicate that users do express emotions of different kinds in the Twitter messages and rich information can be obtain to construct the emotion-based user profiles.

The second research question to be investigated is how different types of emotions impact the variety of emotion-based user profiles. To answer this question, we apply entropy to quantify the information embodied in a emotion-based user profile $P_(u,e)$. The entropy of a given emotion-based user profile is computed as follows.

$$entropy(P(u,e)) = \sum_{c_e \in C_e} p(c_e) \cdot (-log_2(p(c_e))) \tag{4.5}$$

where $p(c_e)$ denotes the probability that the concept c was utilized by the corresponding user and can be modeled via the weight $w(u,c_e)$ of topic c_e in the emotion-based user profile. And using base 2 for the computation of the logarithm can measure entropy in bits.

topic	**user profile** *A*	**user profile** *B*	**user profile** *C*
iPad	4	0	2
Garageband	4	0	4
Apple	4	0	5
French Open	0	6	4
Italy	0	3	1
tennis	0	2	4
entropy	1.06	0.99	2.44

Table 4.5: Entropy of example user profiles.

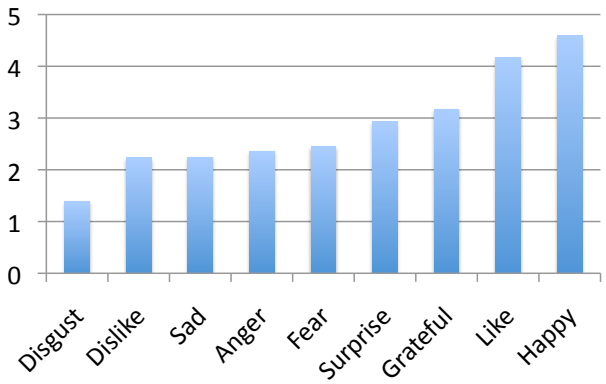


Figure 4.9: The average entropy of emotion-based user profiles for different types of emotions

To clarify the meaning of entropy in context of the user profiles, we calculated the entropy of three example user profiles (see Table 4.5). The entropy of the example profiles depend on the number of topics that appear in the profiles and the corresponding frequencies. The user profile *A* and user profile *B* both contain three distinct topics. However, the entropy of user profiles *A* is higher than the entropy of user profile *B*, in which the topics appear with different probabilities (e.g., $p(FrenchOpen) = 6/11$, $p(Italy) = 3/11$, and $p(tennis) = 2/11$) instead of being uniformly distributed as in user profile *A* (e.g., $p(iPad) = 4/12$, $p(Garageband) = 4/12$, and $p(Apple) = 4/12$). The user profile *C*, which covers a rich variety of topics, has the highest entropy.

For each type of emotion, we take the average of entropy of the corresponding emotion-based user profiles (see Figure 4.9). The emotion-based user profiles constructed for emotion *happy* bears the highest average entropy (4.59 bits). Note that 42.16% of the Twitter messages in our dataset are labeled as *happy*, more than any other types of emotions (see Table 4.4). We further observed that the average

Research question	Summary of findings
How can we correlate microposts with external Web resources?	▶ The exploitation of URLs in well as content of microposts allows for relating microposts with with external Web resources.
How can we exploit emotions expressed in microposts for semantic enrichment?	▶ We succeed in labeling emotions expressed in microposts by utilizing conventional markers.
Which strategy allows for the highest accuracy for (i) correlating microposts with Web resources, and (ii) inferring emotions in microposts?	▶ The entity-based strategy with temporal constraints achieves the highest accuracy for the linkage discovery. ▶ The entity-, and Wordnet-based strategies outperform other strategies in the emotion classification experiments.
How does the semantic enrichment impact the characteristics of user profiles?	▶ The exploitation of external resources allows for constructing more meaningful user profiles. ▶ The emotion-based user profiles allows for better understanding users' opinion about various topics.

Table 4.6: Overview on research questions investigated in this chapter.

entropy of user profiles for the positive emotions (e.g, *happy*, *like* and *grateful*) is higher the negative emotions (e.g, *anger*, *sad dislike*). For example, while the average entropy for emotion *happy* is 4.12 bits which allows for describing 17 possible states, for emotion *dislike* it is only 2.22 bit which could describe 4 possible states. Thus, the results indicate that the variety of topics in positive user profiles is higher than in negative user profiles.

4.4 Discussion

To answer the research questions raised at the begining of this chapter, we introduced and evaluated strategies for enriching the semantics of individual microblogging activities. Further, we analyzed the impact of the semantic enrichment on the quality of user profiles constructed based on microblogging data. We summarize our findings in Table 4.6.

We presented different strategies that connect microposts with related Web resources and exploit semantics extracted from external Web resources to deduce and contextualize the semantic meaning of individual Twitter posts. We evaluated our strategies for linkage discovery in a context of relating tweets to online news articles. Our evaluation on a large Twitter dataset (more than 3 million tweets posted by more than 45,000 users) showed that, given the name of entities mentioned in a news article (such as persons or organizations) as well as the temporal context of the article, we can relate tweets and news articles with high precision (more than 70%) and high coverage (approx. 15% of the tweets can be linked to news articles).

In Chapter 3 we hypothesized that *the exploitation of external Web resources creates more valuable user profiles than micropost-only-based strategies* (see *H2* in table 3.6).

The analysis conducted in this chapter revealed that the exploitation of tweet-news relation has significant impact on user modeling and allows for the construction of more meaningful profiles (more profile facets and more detailed knowledge regarding user interests/concerns) than user modeling based on tweets only. Semantic enrichment of Twitter user activities based on semantics extracted from news articles thus leads to meaningful representations of Twitter activities, ready for being applied in Twitter and other Social Web systems. In Chapter 6, we will further deepen the investigation of how the profiles constructed by this type of user modeling strategies impact personalization on the Social Web. Given the variety and recency of the constructed profiles, there are different applications worthwhile to explore such as Twitter stream and message recommendations, product recommendations or recommending news.

We further exploited the human emotions (e.g., Happy and Sad) expressed in the microposts to enrich the semantics of microblogging activities. We proposed an approach for incorporating human emotions into the construction of user profiles based on microblogging data, which allows us to model individual users' preferences for different emotion facets. By utilizing conventional markers such as emoticons and hashtags, we automatically labeled a large Twitter dataset, which consists of more than 4 million tweets, with a set of fine-grained emotion classes. We conducted emotion classification experiments based on the labeled dataset and applied various topic modeling strategies to extract features from microblogging activities for training the classifier. We investigated the impact of different feature extraction strategies on the performance of emotion classification. The results revealed that the entity-based and Wordnet-based strategies, which extract rich semantics from microblogging streams, clearly outperformed the hashtag-based strategy in the classification experiments. Furthermore, by analyzing the characteristics of user profiles constructed for different emotion facets, we observed that users express more positive sentiment (e.g., Like and Happy) than negative sentiment (e.g., Angry and Fear) in their microposts. The results of our analysis also indicated that the emotion-based user profiles differ between different emotion facets. For example, the variety of positive user profiles is higher than negative profiles. We conclude that the exploitation of emotions expressed in the microposts allows for better grasping individual users' opinions about various topics discussed on the microblogging platforms and therefor offers further semantic enrichment of microblogging activities for user modeling. In Chapter 5, we will conduct sentiment analysis across different platforms and languages and further investigate that how sentiment analysis can be applied to analyze the microblogging behavior for users in different cultural groups.

Chapter 5

Microblogging-based User Modeling for Culture-aware Analytics

We have presented a microblogging-based user modeling framework that provides various strategies for modeling users' preferences (see Chapter 3) and explored methods for enriching the semantics of microblogging activities (see Chapter 4). Given a wide range of design choices featured in our user modeling framework, the user modeling functionality can be adapted to given circumstances. In this chapter, we analyze and compare the microblogging behavior between users from different cultural groups. Such comparative study deliver insights for culture-aware analytics and therefore allows us to adapt the user modeling functionality to a given cultural group. The main contributions of this chapter has been published in [68, 69].

5.1 Introduction

Microblogging services allow people to publish, share and discuss short messages on the Web. Users are enabled to post messages about their daily activities, subscribe to message feeds of other users (following) and propagate individual messages to their followers (reposting). By analyzing individual microblogging activities, it is possible to learn about the characteristics, preferences and concerns of users. Java et al. presented an early attempt to understand the users intentions in using microblogging services such as Twitter by analyzing the topological and geographical properties of Twitter's social network [96]. The authors found four main types of user intentions on Twitter: daily chatter, conversations, sharing in-

formation and reporting news. Macskassy and Michelson analyzed user behavior on Twitter and discovered that users often repost (retweet) messages about topics which are complementary to the topics about which they themselves publish microposts [120]. Furthermore, Naveed et al. discovered that the sentiment expressed in messages is an important feature for predicting whether a message will get reposted or not [120].

Many research efforts have been done, which solely focus on Twitter. Yet, there exists little research on comparing the user behavior across different microblogging platforms. In China, Sina Weibo[1], on which about 100 million messages are posted each day by more than 300 million users[2], is leading the microblogging market since Twitter is unavailable. Both Sina Weibo and Twitter basically feature the same functionality. For example, both services limit the lengths of microposts to 140 characters and allow users to organize themselves in a follower-followee network, where people follow the message updates of other users (unidirectional relationship). Sina Weibo and Twitter provide realtime access to the microposts via APIs and therefore allow for investigating and analyzing interesting applications and functionality such as recommending popular URLs in Twitter [44] or analyzing trending events on Sina Weibo [118, 151]. Yu et al. compare popular trending topics on Sina Weibo with those on Twitter and discover that Twitter trends correlate strongly with news topics while trends on Sina Weibo are typically related to amusement (e.g. links to funny videos or stories) [187]. However, Yu et al. only investigate global trends and do not study individual user behavior. In this chapter, we close this gap and compare users' microblogging behavior on Sina Weibo and Twitter. Based on the user modeling framework and semantic enrichment techniques introduced in the previous chapters, we conduct—to the best of our knowledge—the first comparative study of the microblogging behavior on Sina Weibo and Twitter and relate our findings to theories about cultural stereotypes developed in social science.

People's social behavior is influenced by their cultural values. Cultural differences across countries have been studied in social science research. Geert Hofstede conducted a large-scale quantitative study in 74 countries and developed five dimensions, which are power distance, individualism, uncertainty avoidance, masculinity, and long term orientation, to characterize national culture [91]. Researchers in computer science have applied Hofstede's cultural dimensions on different topics such as studying knowledge sharing in virtual communities [11] or generating adaptive user interface [154]. Research on cultural characteristics of user behavior on the Social Web has also been initiated. For example, Mandl investigates how blog pages, especially the communication patterns between bloggers and commentators, from

[1] http://www.weibo.com/

[2] http://en.wikipedia.org/wiki/Sina_Weibo

China differ from the ones from Germany [121]. Mandl correlates his findings to cultural dimensions proposed by Geert Hofstede. Chen et al. analyze the tagging behavior of two user groups from two popular social music sites in China and Europe respectively and observe differences between the two cultural groups, e.g. Chinese users have a smaller tendency to apply subjective tags but prefer the usage of factual tags [45]. A recent study investigates the influence of national culture on people's online scheduling behavior [156]. The results reveal strong correlation between national culture and people's scheduling behavioral patterns such as that participants from collectivist countries (e.g. China, Japan) respond the polls faster and reach more consensus than highly individualistic countries (e.g. US, the Netherlands). So far, there exists little knowledge about the differences and commonalities regarding the microblogging behavior of users from different cultural groups. In this chapter, we conduct a comparative study of microblogging behavior between Chinese and Western users. We apply techniques including semantic analysis and sentiment analysis that allows us to investigate not only the meaning of the microposts but also the users' opinions revealed in both Chinese and English microposts. Furthermore, we apply temporal analysis to analyze and compare how users' microblogging behavior changes over time on Sina Weibo and Twitter.

The main research questions that we will answer in this chapter are:

- How does the microblogging behavior vary between different cultural groups (e.g. Chinese and American users)?
- Do differences in users' microblogging behavior correlate with cultural theories in social sciences?

In Section 5.2, we will conduct a large scale comparative study of users' microblogging behavior on Sina Weibo and Twitter and reveal the key differences in Western and Chinese microblogging practices, which reflect underlying cultural differences. In Secion 5.3, we will further study the information propagation patterns on the two microblogging platforms by examining users' reposting behavior. We conclude this chapter with a summary of findings and a discussion about the correlation between the findings and cultural models from social science research.

5.2 Analysis of Users' Microblogging Behavior on Sina Weibo and Twitter

In this section, we study user behavior on two different microblogging platforms: Sina Weibo and Twitter. The main contributions of this section can be summarized as follows:

- With functionality for sentiment analysis and semantic enrichment of Chinese microblog posts, we apply our framework for user modeling based on usage data from two microblogging services.
- We conduct intensive analyses based on more than 40 million microblog posts and compare the microblogging behavior on Sina Weibo and Twitter regarding five dimensions: (i) access behavior, (ii) syntactic content analysis, (iii) semantic content analysis, (iv) sentiment analysis, (v) temporal behavior.
- We relate our findings to theories about cultural stereotypes developed in social sciences and therefore explain how our insights can allow for culture-aware user modeling based on microblogging streams.

5.2.1 Methodology

We first detail our research questions and present our user modeling environment that allows us to investigate the research questions.

Research Questions

Our research goal is to study user behavior on Sina Weibo and Twitter to gain insights for user modeling on microblogging streams. Therefore, we investigate (1) how people access microblogging services, (2) the content, (3) semantics, (4) sentiment of microblog posts, and (5) the temporal behavior of users' microblogging activities.

Analysis of Access Behavior Microblogging services such as Sina Weibo and Twitter can be accessed via different client applications from both mobile devices and desktop devices. User behavior that can be observed on a microblogging service may be influenced by the client application and device via which a user accesses the service. We thus first study the following research questions:

- *RQ1.1: How do people access Sina Weibo and Twitter respectively to publish microposts?*
- *RQ1.2: To what extent do individual users access a microblogging service from different client applications?*

Syntactic Content Analysis Both Sina Weibo and Twitter limit the length of posts to 140 characters. This limitation impacts the writing style of microblog users

and may result in characteristic usage patterns. These patterns may also differ between Sina Weibo, where the main language is Chinese, and Twitter, where a large fraction of Twitter messages is posted in English. Regarding the syntactic content analysis, we particularly analyze the usage of hashtags and URLs and answer the following questions:

- *RQ1.3: How does the usage of hashtags, URLs and other syntactic patterns (e.g. punctuation) differ between Sina Weibo and Twitter for both (i) the entire user population and (ii) individual users?*
- *RQ1.4: To what extent is the usage of hashtags and URLs influenced by the users' access behavior?*

Semantic Content Analysis To better understand the semantics of the content that users post on microblogging services, we enhance and apply our user modeling framework to extract semantically meaningful concepts from the microposts published on Sina Weibo and Twitter. In the context of the semantic content analysis, we investigate the following aspects:

- *RQ1.5: What kind of topics and concepts do users mention and discuss on Sina Weibo and Twitter respectively?*
- *RQ1.6: To what extent do the types of concepts that users mention in their posts depend on the client applications via which they publish their posts?*

Sentiment Analysis Microblogs allow users to express and discuss their opinions about topics that people are concerned with. To analyze the sentiment that people reveal in their microblog posts, we also extend our user modeling framework with a sentiment analysis component for English and Chinese messages. We therefore analyze the sentiment of Chinese and English messages and study the following questions:

- *RQ1.7: To what extent do users reveal their sentiment on Sina Weibo and Twitter respectively?*
- *RQ1.8: To what extent does the sentiment correlate with the type of topics and concepts that people mention in their Sina Weibo and Twitter messages?*

Analysis of Temporal Behavior The users' microblogging behavior may change over time and may, for example, differ between working hours and leisure time. Therefore, we investigate the following research questions:

- *RQ1.9: How does the posting behavior of users, particularly regarding the type of topics that the users mention, change between weekdays and weekends on Sina Weibo and Twitter?*
- *RQ1.10: How do individual user interests change over time in the two microblogging services?*

Applying User Modeling Framework for Cultural-aware Analytics

In Chapter 3, we presented a Twitter-based user modeling framework for inferring user interest from tweets. Our framework monitors Twitter activities of a user and enriches the semantics of her Twitter messages by extracting meaningful concepts and topics (e.g. DBpedia concepts) from the messages' content and by linking posts to external relevant Web resources such new articles. Different weighting schemes such as time-sensitive or term-frequency-based functions allow for estimating to what extent a user might be interested in a given concept at a particular point in time. The generated user profiles can therefore be considered as a set of weighted semantic concepts.

We also developed GeniUS which is a software library implemented based on our user modeling framework. GeniUS consists of four main modules including the *Item Fetcher*, *Topic Modeling & Enrichment*, *Weighting Function* and *RDF Serialization* (see Section 3.3). The GeniUS modules are exposed as JAVA interfaces so that news functions can be easily implemented for given circumstances.

In this chapter, in order to collect usage data from different microblogging services and further conduct analyses based on both Chinese and English microposts, we implement three new functions for (1) monitoring microblogging activities and collecting microposts published on Sina Weibo (*Item Fetcher*), (2) extracting topics of interests in Chinese microposts (*Topic Modeling*) and (3) identifying sentiment expressed in both Chinese and English microposts (*Enrichment*). We use ICTCALS[3] as part-of-speech tagger for Chinese text and extract named entities such as locations, organizations and persons from Chinese posts. Further, we extract same types of named entities from English post using OpenCalais. Previous study reported that ICTCALS and OpenCalais achieved close performance for extracting named entities from text, 0.8732 [189] and 0.8793 [162] in precision respectively. To ensure a fair comparison between datasets in Chinese and English, we utilize a collection of common emoticons and affective words in both languages to automatically label the positive and negative sentiment expressed in the microposts. Given these additional features, we are able to apply the same user modeling techniques on both microblogging services Sina Weibo and Twitter and can therefore analyze

[3] http://ictclas.org/

	Sina Weibo	Twitter
number of posts	22,708,173	24,227,492
number of users	6,837,988	1,046,222

Table 5.1: Overview of datasets for Sina Weibo and Twitter.

characteristics and behavior on the Asian and Western microblogging platforms. We tested statistical significance of our results with a two-tailed t-Test where the significance level was set to $\alpha = 0.01$ unless otherwise noted.

Data Collection

Given the framework, we collected microposts over a period of more than two months via the Sina Weibo Open API and the Twitter Streaming API respectively. For Twitter, we started from a seed set of 56 Twitter users and then we gradually extended this set in a snowball manner. Overall, we collected more than 24 million tweets published by more than 1 million users. For Sina Weibo, since it does not provide functionality similar to Twitter's Streaming API, we monitored the most recent public microposts and finally collected more than 22 million microposts published by more than 6 million users (see Table 5.1). Twitter posts and Sina Weibo posts were then processed by our framework in order to enrich the semantics of the posts (e.g. entity extraction, sentiment analysis). To better understand the behavior on the level of individual users, we extracted a sample of 1200 active Twitter users (who post in English) and 2616 active Sina Weibo users. The majority of the Twitter users (more than 80%) is – according to their Twitter profile – from the United States while the great majority of the Sina Weibo users (more than 95%) is located in China.

Based on the more than 40 million posts that we collected from Sina Weibo and Twitter and processed with our user modeling framework, we study the users' behavior on the two platforms and answer the research questions regarding the five dimensions ranging from access behavior to temporal behavior.

5.2.2 Analysis of Access Behavior

Results

We first analyzed the most popular client applications that people use to publish posts on Sina Weibo and Twitter. On both platforms, the Web interface is the most

type of access	fraction of posts Weibo	Twitter
posted on a Web or desktop application	54.9	66.2
posted on a mobile application	45.1	33.8
primary product of microblogging activity	90.6	96.7
byproduct of an activity on another platform	9.4	3.3

Table 5.2: Fraction of posts for different categories of clients

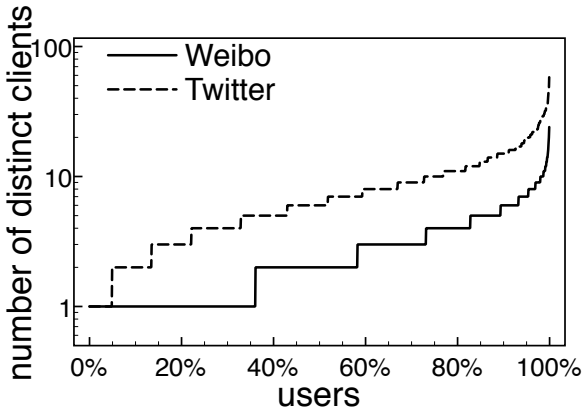


Figure 5.1: Number of distinct access clients for individual users

popular way to access the microblogging services: 43.1% of the posts are published via the Web on Sina Weibo and 38.5% on Twitter. Other popular clients on Sina Weibo are mainly designed for mobile devices such as the iPhone (7.6%) and Nokia devices (9.4%). Among the most popular Twitter clients are many desktop-based applications such as *TweetDeck*, via which 10.7% of the posts are published. Moreover, we observe on both platforms that people publish posts that are rather byproducts of activities the users perform on other platforms. For example, 1.3% of the posts in our Twitter dataset are published via *Twitterfeed*, an application that allows for publishing announcements on a user's Twitter timeline whenever she publishes a new blog article.

In Table 5.2, we overview the type of client applications that people use to publish microblog posts. We therefore manually categorized the 50 most popular clients, that generate more than 90% of the posts on both microblogging services.

We observe that the fraction of posts that are published via mobile devices is significantly higher on Sina Weibo (45.1%) in comparison to Twitter (33.8%). Furthermore, we discover that the fraction of posts which are rather byproducts of other Web activities of the users – hence where the intent of the actual user activity was not targeted towards Sina Weibo or Twitter – is almost three times higher on Sina Weibo than on Twitter ($p < 0.01$).

In Figure 5.1, we plot for each of the sample users the number of distinct applications which they utilize for publishing microposts[4]. We see that on Twitter more than 95% of the people use more than one client application while on Sina Weibo around 65% of the users switch between different clients.

Findings

From the results above, we conclude the analysis of access behavior with two main findings, referring to the research questions *RQ1.1* and *RQ1.2*:

- *F1:* On both platforms, the major way to accessing the microblogging services is via the official Web interfaces or desktop-based applications. Chinese users seem to differ from the English-spoken Twitter users regarding two core aspects: (i) they use mobile applications more extensively and (ii) publish microposts more often as a byproduct of their other Social Web activities.
- *F2:* The results regarding the individual users' access behavior illustrate that Twitter users switch between different clients to access the microblogging service more often than the users on Sina Weibo. This difference in behavior could be explained by the lower overall number of valuable Sina Weibo client applications (e.g. in our dataset: 3015 different Sina Weibo clients versus 5468 Twitter clients).

5.2.3 Syntactic Content Analysis

Results

In Table 5.3, we compare the syntax of messages posted on Sina Weibo and Twitter and particularly the usage of hashtags and URLs. Overall, 20% of the Twitter messages contain hashtags and 29.1% of the tweets feature a URL. Therefore, the usage of hashtags and URLs on Twitter is 3.2 times and 1.97 times respectively more intensive than on Sina Weibo ($p < 0.01$). The analysis of special characters

[4]Note that the x-axes of the diagrams refer to user percentiles.

syntactic characteristics posts that contain:	proportion of posts Weibo	 Twitter
hashtags	6.3%	20.0%
URLs	14.8%	29.1%
question marks "?"	9.9%	18.6%
exclamation marks "!"	26.1%	20.7%
"?" and "!"	3.1%	3.5%

Table 5.3: Comparison of syntactic content analysis

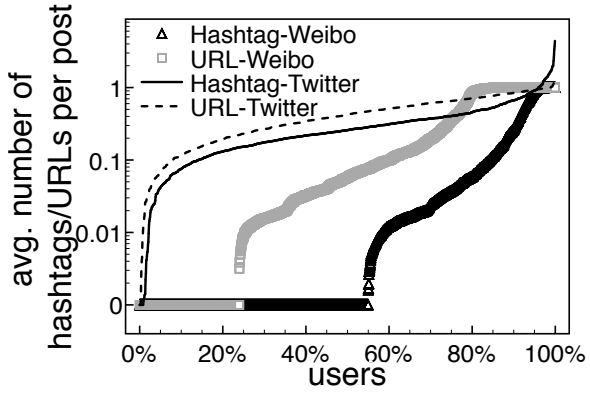


Figure 5.2: Comparison of writing style for individual users

implies that users on Twitter ask more than twice as many questions than users on Sina Weibo (see question marks in Table 5.3). In contrast, Sina Weibo users make more extensive use of exclamation marks and therefore more often put extra emphasis on their statements. On both platforms, we observe that positive emoticons outrange negative emoticons which indicates that people are more likely to make positive statements (cf. Section 5.2.5).

To further analyze the usage of hashtags and URLs, we also plot for each individual user in our samples the average number of hashtags and URLs per post. From Figure 5.2, we infer that a considerably high fraction of Sina Weibo users does not mention hashtags or URLs at all. For 55% of the Chinese microbloggers on Sina Weibo, we did not observe any hashtag. In contrast, on Twitter the people make more frequently use of hashtags or URLs. For example, for more than 85% of the Twitter users, the average number of hashtags per post is at least 0.1, i.e. at least every tenth micropost mentions a hashtag, and 3.9% of the users mention, on average, even more than one hashtag per tweet.

In Table 5.4 we analyze the influence of the access behavior (see Sect. 5.2.2) on

posts contain	proportion of posts			
	Weibo		Twitter	
	Desktop/Mobile	*Microblog/Byproduct*	*Desktop/Mobile*	*Microblog/Byproduct*
hashtags	6.5%/3.5%	3.8%/17.9%	20.7%/18.6%	19.9%/21.3%
URLs	17.8%/5.2%	5.7%/73.5%	31.6%/20.1%	25.3%/97.9%

Table 5.4: Impact of the access behavior on the syntactic characteristics of microposts.

the usage of hashtags and URLs. For both services, we observe that the usage of hashtags and URLs decreases slightly when people publish microposts from their mobile devices instead of their desktop computers. This difference is more significant on Sina Weibo. For example, on Sina Weibo the number of posts that contain a URL and are issued from a desktop application (17.8%) is more than three times higher than the one for mobile devices (5.2%). On Twitter, the usage of URLs on desktop devices is only 1.57 times higher than on mobile devices. Regarding the type of activity that a user performed to publish a micropost, we observe that 97.9% of the tweets that were generated as byproducts of other activities (e.g. publishing an article in a blog or "check-in" activities on *Foursquare*) contain URLs. In contrast, for the conventional microblogging, only 25.3% of the Twitter messages contain URLs. A similar increase can be observed on Sina Weibo. The number of hashtags is slightly less influenced by the type of activity that caused a micropost (see Table 5.4). In particular for Twitter, the increase in the number of messages that contain a hashtag is with less than 8% rather low.

Findings

Given the results above, we can answer *RQ1.3* and *RQ1.4* as follows:

- *F3:* Overall, the results show that hashtags and URLs are less frequently applied on Sina Weibo than on Twitter. This finding holds for both (i) the entire user population and (ii) individual users. In fact, we observe that a large fraction of users on Sina Weibo does not make use of hashtags which implies that hashtag-based user profiles, as discussed in [5], or topic modeling based on hashtags, as proposed by Romero et al. [163] do not seem to be appropriate on Sina Weibo. The usage statistics regarding question marks indicate that Twitter users ask more questions than Sina Weibo users.
- *F4:* The usage of hashtags and URLs is moreover influenced by the access behavior. We discover that (i) users are more likely to use hashtags and URLs

when they post messages via desktop applications than via mobile applications. Furthermore, (ii) whenever messages are published as a byproduct of another activity – where the primary intention of the user is rather the promotion of an activity that the user performed on another platform – the probability that a micropost contains a hashtag or URL increases. A large fraction of these *byproduct microposts* seems to be automatically generated based on the activity the user performed on another platform. For user modeling those posts offer means to further contextualize the microblogging activities by following the URLs that are contained in the posts (cf. [5]).

5.2.4 Semantic Content Analysis

Results

Based on the semantic enrichment provided by our user modeling framework, we analyze and compare the types of concepts and topics that people mention in their microposts on Sina Weibo and Twitter respectively. In Table 5.5 we compare the usage of three types of entities (location, people and organization). Most of the extracted semantic concepts refer to locations (e.g. cities, points of interests): 58.4% for Sina Weibo and 44.6% for Twitter. On Twitter, posts that refer to organizations (e.g. companies, institutions) are more than four times more likely to appear than on Sina Weibo. Examples of entities that were trending on Twitter include different types of entities such as “Mubarak” (person), the former president of Egypt, or “Republican Party” (organization). In contrast, the most popular entities on Sina Weibo are related to locations such as “Beijing” or “United States”.

Figure 5.3 depicts the average number of entities that can be extracted per post for the individual users in our sample. For 24.8% of the Sina Weibo users, one can detect, on average, more than one entity per post. Moreover, the fraction of users for whom no entity can be extracted is 7.9% in contrast to 10.1% on Twitter. The semantics of the users’ messages posted on Sina Weibo are therefore easier to deduce than on Twitter. Based on a comparison of a sample of individual Chinese and English microposts, we hypothesize that this is caused by the expressivity of the Chinese language: while Twitter users are often forced to leave out entities or use abbreviations to refer to entities, Sina Weibo users can exploit the 140 characters more effectively.

Table 5.5 illustrates how the access behavior influences the semantics of the microposts. When users publish posts from their mobile devices, then it becomes less likely, in comparison to access via desktop (tailored Web) applications, that a message mentions an entity. For microposts that are byproducts of other Web

	proportion of posts	
types of posts	**Weibo**	**Twitter**
Location	58.4%	44.6%
Organization	3.3%	16.0%
Person	38.3%	39.4%

Impact of the access behavior on the type of concepts mentioned in the posts

	Desktop/ Mobile	*Microblog/ Byproduct*	*Desktop/ Mobile*	*Microblog/ Byproduct*
Location	11.2%/6.6%	15.5%/4.0%	9.3%/8.4%	8.9%/13.7%
Organization	0.7%/0.6%	0.9%/0.4%	3.5%/2.9%	3.3%/4.5%
Person	12.4%/12.3%	17.4%/4.9%	8.1%/6.7%	7.6%/8.7%

Table 5.5: Semantic analysis overall and impact of access behavior on the semantics.

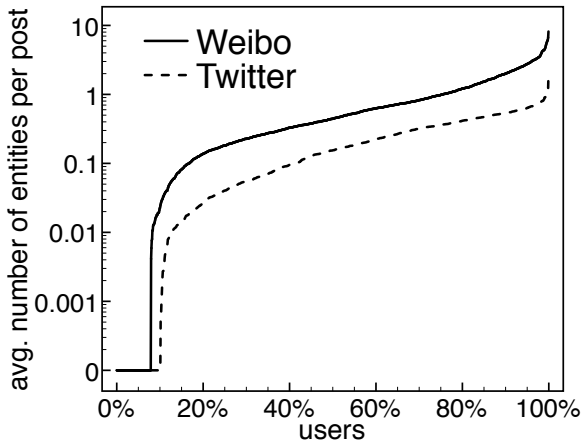


Figure 5.3: Semantic analysis for individual users

activities (e.g. activities on *Foursquare*), we observe that it becomes more likely that entities and particularly location entities are mentioned in a post on Twitter. In contrast, on Sina Weibo users mention more entities in context of their standard microblogging activities.

Findings

The results of the analysis illustrate the commonalities and differences regarding the semantic meaning of the microposts that users publish on Sina Weibo and Twitter

type of posts	proportion of positive/negative posts Weibo	Twitter
Overall	78.8%/21.2%	70.5%/29.5%
posts that mention certain types of entities:		
Location	82.7%/17.3%	65.6%/34.4%
Organization	78.5%/21.5%	70.1%/29.9%
Person	82.8%/17.2%	65.7%/34.3%

Table 5.6: Sentiment expressed in overall posts and posts that mention certain types of topics

respectively (see *RQ1.5* and *RQ1.6* in Section 5.2.1):

- *F5:* The topics that users discuss on Sina Weibo are to a large extent related to locations and persons. In contrast to Twitter, users on Sina Weibo avoid talking about organizations such as political parties or other institutions. Overall, the semantics of Sina Weibo messages can be better extracted than the semantics of tweets. Consequently, when modeling the microblogging activities for individual users, entity-based user profiles [5] can more successfully be generated for Sina Weibo users: for 92.1% of them one can identify at least one entity of interest in comparison to 89.9% on Twitter.
- *F6:* The type of applications via which users access the microblogging services, affects the occurrence of semantic concepts in the microposts. On mobile devices people tend to mention less entities than on desktop devices. Furthermore, microposts on Twitter are more likely to mention entities and locations particularly if the post was generated as a byproduct of an activity performed on another platform.

5.2.5 Sentiment Analysis

Results

The sentiment analysis provided by our framework classifies microblog posts as either positive, negative or neutral. Overall, 83.4% and 82.4% of the Sina Weibo and Twitter posts respectively were classified as neutral. Table 5.6 overviews the sentiment polarities of those posts that have been classified as positive or negative. On Sina Weibo the portion of positive posts (78.8%) is clearly higher than on Twitter (70.5%). In Figure 5.4 we plot the ratio of positive posts with respect to all posts,

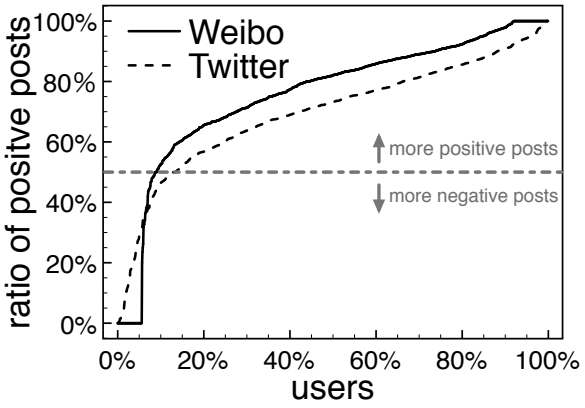


Figure 5.4: The ration of positive posts on two microblogging services

which either have a positive or negative sentiment, for individual users: 92.5% of the users publish more positive messages than negative ones on Sina Weibo in comparison to 86.4% for the Twitter users. On Sina Weibo, we also discover a considerable fraction of users for whom the non-neutral posts are always positive (8.0%) or always negative (5.6%).

In Table 5.6 we moreover analyze the sentiment revealed in the microposts that mention certain types of entities. Again, the proportion of positive posts exceeds the proportion of negative posts clearly and Sina Weibo users tend to be more positive towards mentioned entities than Twitter users. Interestingly, whenever locations or persons are mentioned in Sina Weibo messages then the likelihood that the post is positive increases on Sina Weibo (from 78.8% to 82.7% and 82.8% respectively) while on Twitter the opposite can be observed (decrease from 70.5% to 65.6% and 65.7% respectively).

Findings

Regarding the research questions *RQ1.7* and *RQ1.8* about the sentiment that users express in their microposts, we conclude the following:

- *F7:* We observe that on both platforms there are significantly more positive posts than negative ones. Moreover, users on Sina Weibo have a stronger tendency to publish positive messages than Twitter users. In fact, the probability for positive messages is 11.8% higher on Sina Weibo than on Twitter ($p < 0.01$).

- *F8:* The sentiment that is expressed in microposts correlates with the type of

posts per weekend day / posts per weekday	Weibo	Twitter
Overall posts	1.19	0.89
posts that mention certain types of entities:		
Location	0.81	1.05
Organization	1.50	0.91
Person	1.19	0.97

Table 5.7: Ratio between weekend posts and weekday posts

concepts that are mentioned in the posts. On Sina Weibo posts that mention locations or persons are more likely to be positive than posts containing organizations. While on Twitter, the opposite can be observed: people talk more positively about organizations than about persons or locations.

5.2.6 Analysis of Temporal Behavior

Results

In Table 5.7 we first compare the posting behavior of users between working days and weekend days by calculating the ratio between the average number of posts per day published during the weekends (Saturday-Sunday) and the one during the week (Monday-Friday). For Sina Weibo this ratio is 1.19, which means that Sina Weibo user publish, on average, 19% more messages per day on the weekend than they do during the week. On the other hand, the users on Twitter publish, on average, 11% less posts during the weekend. Therefore, it seems that microblogging in China has not penetrated the daily (possibly work-related) routines as strongly as it does in Western countries.

In Figure 5.5 we plot the weekend-weekday ratio for the individual users. While the overall amount of microblogging activities per day on Sina Weibo is higher on the weekends than during the day, we also discover that 1.2% of the Sina Weibo users perform microblogging activities solely during the weekend (ratio of weekend posts is infinite). For about 50% of the users on Sina Weibo the weekend-weekday ratio is greater than 1 which means that they publish more frequently during the weekend. In contrast, on Twitter we identify only 28% of the users who publish more tweets per day on a weekend than during a weekday.

As depicted in Table 5.7, the occurrence of organizations and persons is more likely during the weekend than during the week on Sina Weibo whereas locations

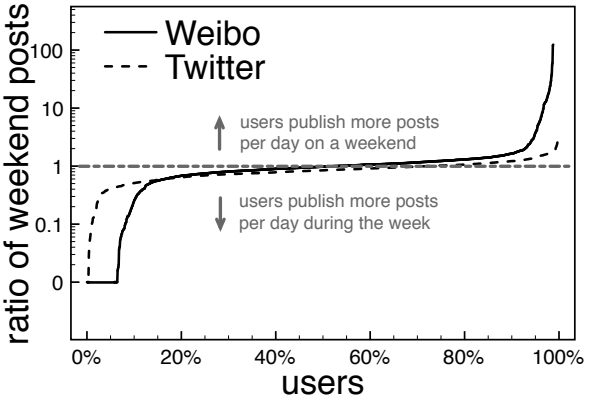
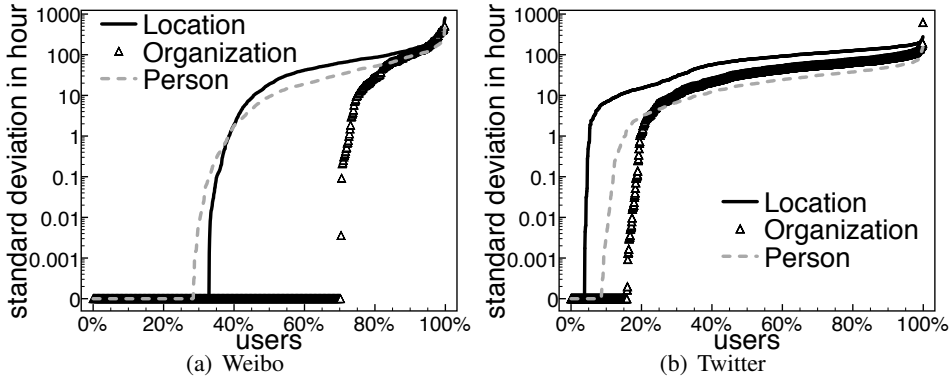


Figure 5.5: Weekend-weekday ratio per user



(a) Weibo (b) Twitter

Figure 5.6: Comparison of topic drift.

appear more likely during a weekday. On Twitter, the opposite characteristics can be observed. For example, Twitter users mention locations more frequently during the weekend than during the week. These differences in mentioning entities during weekends/weekdays on Sina Weibo and Twitter respectively may relate to different life styles that Chinese and Western people follow. Investigating the particular reasons for them can be interesting for future work.

Furthermore, we study how individual user interests change over time by calculating the standard deviation of the timestamps of microposts that mention a certain topic (entity). The higher the standard deviation of a certain topic the longer the time period over which the topic is mentioned in the posts. In Figure 5.6 we plot for each user the average standard deviation of the topics which a user mentioned at least once, and group the average standard deviations by the type of the topics. Overall, we observe that topics on Sina Weibo seem to fluctuate stronger than on

	China	US
Power distance	80	40
Individualism	20	91
Masculinity	66	62
Uncertainty avoidance	40	46
Long term orientation	118	29

Table 5.8: Hofstede's cultural index for China and United States

Twitter. Sina Weibo users often mention certain concepts only once. For example, for more than 80% of the Sina Weibo users of our sample, the standard deviation of the organization-related topics is 0. These users mention thus organizations only once in their posts. On both platforms the location-related concepts are, on average, mentioned over a longer period of time than organization-related and person-related concepts.

Findings

The main findings from the analysis of the temporal behavior (research questions *RQ1.9* and *RQ1.10*) can be summarized as follows:

- *F9:* On both platforms, the users posting behavior during weekdays differs the one during weekend: while users on Sina Weibo are more active on the weekends, Twitter users tend to be more active during weekdays. Moreover, user interests change between weekends and weekdays. Again, this change of interests differs between Sina Weibo and Twitter users: while for Sina Weibo users we observe a rising interest in persons and organizations during the weekend, the interests of Twitter users focus more on locations. These findings imply that it is beneficial to adapt user interest profiling to the temporal as well as to the cultural context.

- *F10:* User interests change over time. On Sina Weibo, the user interests seem to have a shorter lifespan than on Twitter. Especially, the individual users interests regarding organization-related topics vanish quickly on Sina Weibo while locations feature the longest span of interests.

5.2.7 Interpretation of Findings

Some of our findings can be explained also by cultural differences between the Chinese Sina Weibo users and the Twitter users who are mainly located in the U.S. (more than 80% of the Twitter sample users are located in the United States). According to Hofstede's cultural index [91], people in China can, for example, be characterized by a higher *power distance* than people from the U.S. (see Table 5.8). This difference might explain our finding *F1* regarding the access behavior (see Section 5.2.2): Sina Weibo users more frequently generate microposts as a byproduct of their other Social Web activities. Therefore, it seems that they are, in comparison to the people who use Twitter, less afraid of disclosing information about themselves. Given the high power distance that is specific to the Chinese culture, we assume that this behavior can be observed because Chinese users do not attribute much impact to their individual activities, i.e. the impact of disclosing information is less because of the high power distance. The more intensive usage of hashtags and URLs which is characteristic for the Twitter users (*F3*, see Section 5.2.3), may relate to both the lower *power distance* and the higher degree of *individualism* of American people (see Table 5.8). By mentioning a hashtag, microbloggers ensure that their message will appear in the public discussions. Twitter users seem to be more eager to let their posts appear in the public discussion. Hence, they seem to have a stronger belief that their post makes a difference (power distance) and possibly also a higher demand to profile themselves in the public discussions (individualism).

We also observed that Sina Weibo users less frequently mention organizations in their posts than Twitter users (*F5*, see Section 5.2.4). This observation is in line with Hofstede's observation that "employee commitment to an organization is low" in China[5], which is one of the typical indicators for a high *long term orientation*. The sentiment analysis (see Section 5.2.5), which showed that the Chinese Sina Weibo users are more positive than the Twitter users from the U.S. (*F7*), further supports this cultural difference regarding the long term orientation. In the context of the sentiment analysis, we furthermore discovered that Sina Weibo users are more positively talking about persons than Twitter users (*F8*) which again supports the Chinese tendency for *collectivism* rather than *individualism*.

The temporal analysis (see Section 5.2.6) revealed that Sina Weibo users are less actively publishing microblog posts during the working days and particularly mention less frequently organizations than during the weekend. This can be interpreted as an indicator for *long term orientation* as it implies a rather low commitment for the organization that the user is working for. Sina Weibo users also seem to change their interests rather quickly in comparison to Twitter users (*F10*). While this seems

[5] http://geert-hofstede.com/china.html

to contradict to the long term orientation of Chinese people, it also reveals that Chinese people adapt faster to new topics which may be interpreted as "an ability to adapt traditions to changed conditions", one of the characteristics of cultures with high long term orientation.

We have examined users' posting behavior on two different microblogging platforms. Our findings reveal significant differences in the microblogging behavior between Chinese and Western users and deliver valuable insights for multilingual and culture-aware user modeling based on microblogging data. In the next section, we will further analyze information propagation and compare the reposting behavior on Sina Weibo and Twitter.

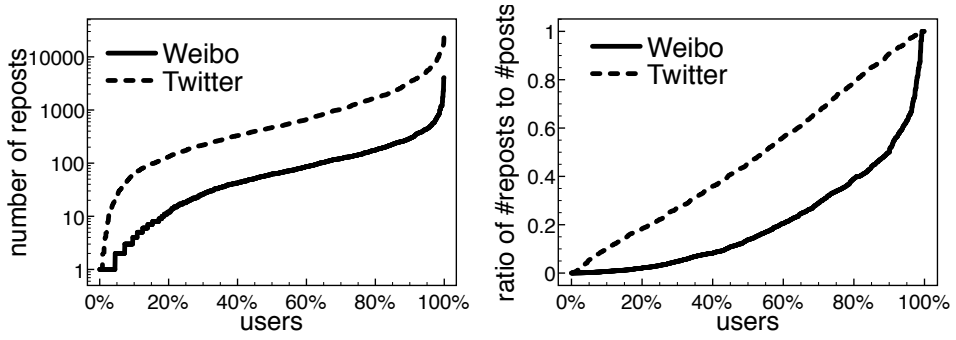
5.3 Analysis of Information Propagation on Sina Weibo and Twitter

Given the microblogging activitites performed by our sample users on Sina Weibo and Twitter, in this section we study their reposting behavior regarding five different perspective: (1) the reposting frequency, (2) the temporal characteristics of the reposting behavior, (3) the broadness of interests of users in reposting content, (4) the content of reposted messages and (5) the sentiment of message that users repost.

5.3.1 Research Questions

Based on the user modeling environment that is capable of processing both Chinese posts from Sina Weibo and English posts from Twitter (see Section 5.2), we investigate the following research questions.

- *RQ2.1: How frequently do users repost messages on Sina Weibo and Twitter respectively?*
- *RQ2.2: How quickly do users propagate information on Sina Weibo and Twitter?*
- *RQ2.3: To what extent does the broadness of user interests vary between Sina Weibo and Twitter?*
- *RQ2.4: What are syntactical characteristics of messages that people propagate on Sina Weibo and Twitter?*
- *RQ2.5: What are the sentiment characteristics of messages that are propagated on Sina Weibo and Twitter?*



(a) #posts (b) repost ratio

Figure 5.7: Overview of reposting behavior for individual users.

We therefore investigate both the actual reposting behavior (RQ1-RQ3) as well as the content of the messages that are propagated (RQ4-RQ5). Some of our findings correlate with cultural characteristics that are, according to Hofstede, attributed to Chinese and American people respectively. We tested statistical significance of our results with a two-tailed t-Test where the significance level was set to $\alpha = 0.01$ unless otherwise noted.

5.3.2 Reposting Frequency

We first overview the reposting behavior on Sina Weibo and Twitter. In Figure 5.7(a), we plot for each of the sample users the number of reposted messages on the two microblogging platforms. It shows that the users on Twitter more frequently repost messages. During the observation period, more than 84% of the sample users on Twitter performed more than 100 reposting activities while on Sina Weibo this number is much lower (36%). To illustrate more clearly how often the users repost messages, we also plot in Figure 5.7(b) for each user the *repost ratio*, i.e. the number of messages reposted by a user divided by the number of all microposts published by this user. On Twitter, we observe a linear distribution. For example, for 80% of the users the repost ratio is less or equal than 0.8. In contrast, on Sina Weibo the distribution of the repost ratio has exponential characteristics. For 80% of the users the repost ratio is less or equal to 0.4.

Referring to *RQ2.1*, we can conclude based on the results above that the reposting behavior differs clearly between the two microblogging platforms: Twitter users perform reposting activities much more frequently than Sina Weibo users.

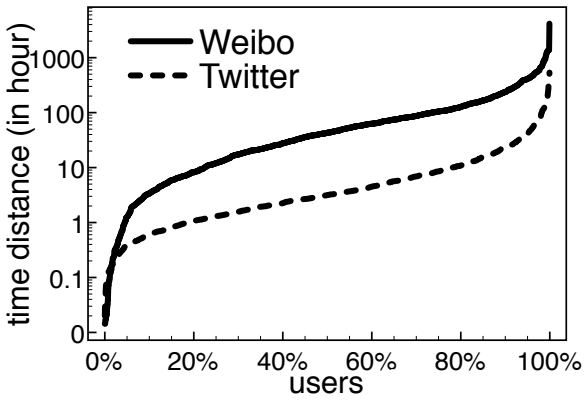


Figure 5.8: How fast do users repost?

5.3.3 Reposting Speed

We further analyze how fast the users repost messages on Sina Weibo and Twitter. For each reposting activity, we calculate the distance between the time when the reposting occurred and the time when the original message was published. In Figure 5.8, we plot the average time distance of all the reposting activities for each user. We observe that, on average, Twitter users repost messages faster than on Sina Weibo. For example, 88% of the Twitter users perform the reposting activity, on average, within the first 24 hours after the original message was published whereas on Sina Weibo only 37% of the users repost a message within the 24 hours. Regarding *RQ2.2* it therefore seems that information propagates more quickly on Twitter than on Sina Weibo. This finding may be explained by the differences regarding the trending topics on the two platforms [187]: Twitter trends are related to news-related information which may change more quickly than the amusement-related information that trends on Sina Weibo. Consequently, users may be triggered to propagate news-related information more quickly than information related to amusement.

5.3.4 Broadness of User Interests

To investigate user interests, we analyze the broadness of information sources from which users propagate messages. In Figure 5.9 we therefore plot for each user the ratio of the number of reposting activities to the number of distinct users who published the original microposts, i.e. the smaller the ratio the higher the broadness of interest. Figure 5.9 depicts that for 90% of the Sina Weibo users the ratio (*repost : reposted users*) is less than 5 which indicates that these users, on average, propagate less than 5 messages from the same source. In contrast, we observe that for only

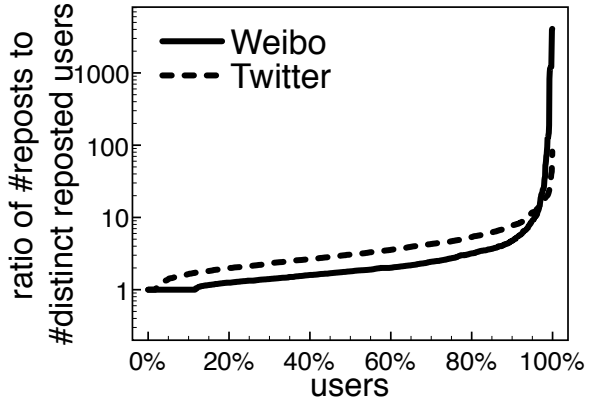
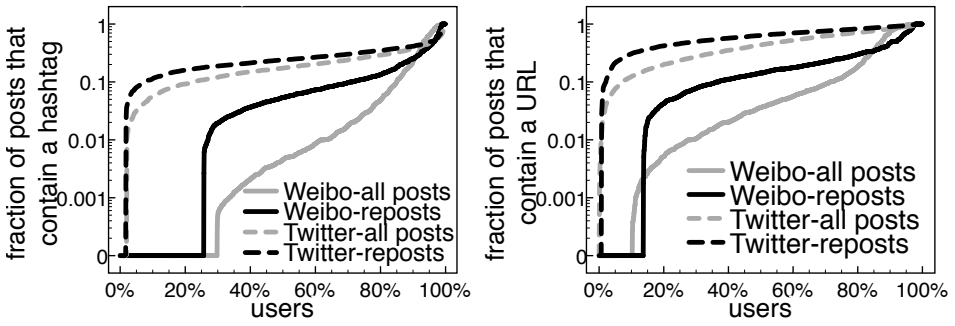


Figure 5.9: Comparison of the interest focus



(a) the usage of hashtags

(b) the usage of URLs

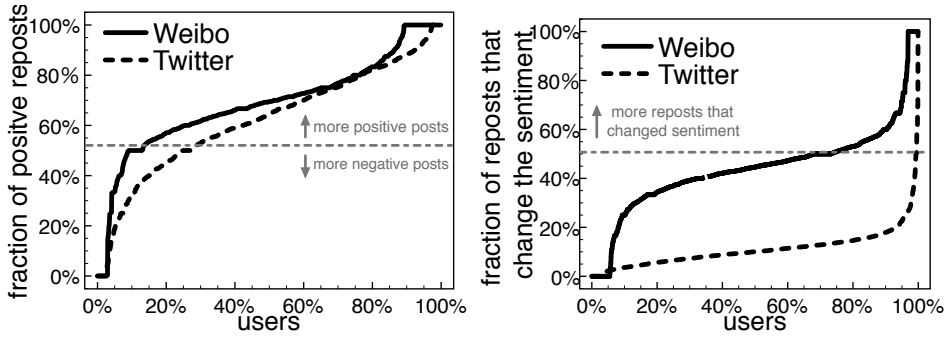
Figure 5.10: Comparison of the syntactic characteristics

76% of the Twitter users the ratio is less than 5 while for the other users the number is higher, thus indicating that these users frequently propagate information from the same sources. Regarding research question *RQ2.3*, we therefore conclude that user interests are broader on Sina Weibo than on Twitter.

5.3.5 Syntactical Characteristics of propagated messages

We compare two important syntactic characteristics of messages that are propagated on Sina Weibo and Twitter: hashtags and URLs. In Figure 5.10, we therefore plot for each user the fraction of reposts that contain hashtags (Figure 5.10(a)) or URLs (Figure 5.10(b)) with respect to the overall number of reposts performed by a user.

On both platforms, reposted messages (black curves in Figure 5.10) are more



(a) fraction of the positive reposting activities (b) changing the sentiment by reposting

Figure 5.11: Sentiment analysis of the reposts

likely to contain hashtags and URLs than other messages (gray curves in Figure 5.10). It thus seems that users are more likely to repost messages that contain hashtags or URLs. For example, for more than 70% of the users the *hashtag ratio* (= fraction of messages that contain hashtags) for reposted messages is higher than 0.01 while only less than 40% of the users feature a hashtag ratio of at least 0.01 for arbitrary messages. On Twitter, users seem to be much more triggered by hashtags and URLs when propagating information. Figure 5.10(b) depicts that for 97% of the users the *URL ratio* (= fraction of messages that contain a URL) is higher than 0.1, i.e. more than 10% of the messages which these users propagate contain a URL. In these situations, the content of the external resource that is linked from a Twitter message might be more important than the actual micropost message. Twitter therefore seems to be more extensively used as a platform for sharing pointers to external Web resources than Sina Weibo while on Sina Weibo the actual discussions and conversations seem to be more predominant than on Twitter.

With respect to research question *RQ2.4*, we can thus conclude that the presence of hashtags and URLs are typical characteristics of messages that are being propagated on both platforms. Hashtags and URLs seem to play an even more important role for Twitter users than for Sina Weibo users when considering information propagation.

5.3.6 Sentiment Characteristics of propagated messages

The sentiment analysis module that we implemented classifies microblog posts as either positive, negative or neutral. Overall, the majority of the messages, which are reposted, is classified as neutral: 75.4% for Sina Weibo and 83.5% for Twit-

ter. In Figure 5.11(a), we plot for each user the ratio of reposted messages with positive sentiment with respect to all reposting activities, which either have a positive or negative sentiment. On Sina Weibo, 91% of the users repost more positive messages than negative ones in comparison to 75% of the users on Twitter. Both microblogging platforms allow users to add comments to a message that they intend to propagate. By adding comments users may change the sentiment of a message (e.g. joking about messages). In Figure 5.11(b) we thus plot the proportion of the reposting activities which change the sentiment of the original message: on Sina Weibo we identify a considerable high fraction of users (32%) for whom more than half of the reposting activities change the sentiment of the original post in comparison to just a few of such users (less than 1%) on Twitter.

Two main findings regarding the sentiment analysis of the reposting behavior can be drawn from the results above to answer research question *RQ2.5*: (1) we observe that the reposting activities are more likely to have positive sentiment on Sina Weibo than on Twitter and (2) users on Sina Weibo change the sentiment of a message which they propagate more often than Twitter users.

5.3.7 Interpretation of Findings

Some of the above findings can also be explained by cultural differences between the Chinese Sina Weibo users and the Twitter users who are mainly located in the U.S. (more than 80% of the Twitter sample users are located in the United States). According to Hofstede's cultural index [91], people in China can, for example, be characterized by a higher *power distance* than people from the U.S. (see Table 5.8). This characteristic may explain our results which indicate that Twitter users more frequently and faster propagate information than Sina Weibo users. Twitter users therefore may have the impression that they play an important role in the information propagation process, i.e. they act as if they are in the power of spreading news.

We also discover that Twitter users have a narrower focus regarding the information streams from which they repost messages while Sina Weibo users select from a broader set of information sources (see Section 5.3.4). We interpret this finding as a signal for *individualism* which is less characteristic for the Chinese culture than for the American culture (see Table 5.8): Chinese microblogging behavior follows a rather collectivistic culture where the actual content of a message seems to be more important than the source which published the content.

The sentiment analysis revealed that Chinese microbloggers have a stronger tendency to propagate positive messages than the merely western microblogging users. The positive nature of the propagated information that people propagate on Sina Weibo might point at the *long term orientation* that is attributed to the Chinese cul-

Analysis	Finding	Cultural dimension
Access	Sina Weibo users more frequently publish messages as a byproduct of their other Social Web activities	Power distance China (high), US (low)
Syntactic	Hashtags and URLs are less frequently applied on Sina Weibo than on Twitter.	Individualism China (low), US (high)
Syntactic	Twitter users are more triggered by hashtags and URLs when propagating information than Sina Weibo users.	Individualism China (low), US (high)
Semantic	Sina Weibo users less frequently mention organizations in their post than Twitter users.	Long term orientation China (high), US (low)
Semantic	User interests are broader on Sina Weibo than on Twitter.	Individualism China (low), US (high)
Sentiment	Sina Weibo users have a stronger tendency to publish positive messages than Twitter users.	Long term orientation China (high), US (low)
Temporal	Twitter users repost messages faster than Sina Weibo users.	Power distance China (high), US (low)
Temporal	Sina Weibo users are less actively pubilshing posts during the working days.	Long term orientation China (high), US (low)

Table 5.9: Some findings and their correlation with cultural dimensions.

ture[6].

5.4 Discussion

In summary, we answer the research questions raised at the beginning of this chapter as follows: (1) different cultural groups show also different microblogging behavior, and (2) these differences can be correlated with theories from social science (see Table 5.9)

Given the framework and semantic enrichment techniques, in this chapter we conducted a large scale analysis of microblogging behavior across different platforms and cultural groups. We analyzed and compared user behavior on two different microblogging platforms: (1) Sina Weibo which is the most popular microblogging service in China and (2) Twitter. Such comparison has not been done before at this scale and is therefore essential for understanding user behavior in the microblogging sphere. In our study, we first analyzed more than 40 million microblogging activities and investigated microblogging behavior from different angles. We (i) analyzed how people access microblogs and (ii) compared the writing style of

[6]http://geert-hofstede.com/china.html

Sina Weibo and Twitter users by analyzing syntactic features of microposts. Based on semantics and sentiments that our user modeling framework extracted from English and Chinese posts, we studied and compared (iii) the topics and (iv) sentiment polarities of posts on Sina Weibo and Twitter. Furthermore, (v) we investigated the temporal dynamics of the microblogging behavior such as the drift of user interests over time.

We further analyzed the reposting behavior in order to study the information propagation cultures and discovered significant differences in the behavior of the Chinese Sina Weibo users and the American Twitter users. For example, Twitter users perform more frequently and faster reposting activities than the users on Sina Weibo. In contrast, Sina Weibo users consider a broader range of information sources from which they repost messages. Moreover, their flavor for propagating messages that have a positive sentiment is more pronounced than for the Twitter users.

Our comparative study reveals that there are significant differences in the users' microblogging behavior between Chinese and Western users on the two microblogging platform. Some of our findings correlate with theories about cultural stereotypes developed in social sciences (see Table 5.9). Therefore, our results are not only of interest for computer science researchers, but also enable social science researchers to confirm their hypotheses about cultural commonalities and differences. Independent from these interpretations of our results, we have given an innovative basis for analyzing microblogging behavior on Sina Weibo and Twitter. Thus, our findings provide valuable insights for culture-aware user modeling and adaptation.

We have demonstrated that how our user modeling functionality can be adapted to a specific application – the cultural-aware analytics. To study the users' microblogging behavior across platforms and cultural groups, we developed new functions that implement modules in our user modeling framework for collecting data from different microblogging services as well as modeling users' interests and behavior patterns based on both Chinese and English microposts. Moreover, the various design dimensions featured in our user modeling framework allows us to explore the differences between Chines and Western users across microblogging platforms from different angels, ranging from the syntactic characteristics in microposts to the temporal patterns that occur in microblogging-based user profiles. Therefore, our framework allows for a comprehensive understanding of users' microblogging behavior which is important for building personalized applications. In the next chapter, we will apply the user modeling framework in various recommender systems and further investigate the impact of different design dimensions on the performance of these recommender systems.

Chapter 6

Microblogging-based User Modeling for Personalized Recommendations

Given the user modeling framework introduced in Chapter 3, we have studied users' microblogging behavior on different microblogging platforms (see Chapter 5). In this chapter, we further analyze and evaluate our microblogging-based user modeling framework in action and apply it in the context of different personalized recommender systems. The main contributions of this chapter have been published in [4, 5, 66, 67].

6.1 Introduction

On microblogging platforms, users are overwhelmed by the massive amount of information available. For example, given the huge amount of information disseminated daily on Twitter, user profiling and personalization that support users in ranking sources to follow [43, 81, 85] or selecting content to read [44, 108, 148] is becoming crucial. Recently, researchers started to explore ranking and recommendations of Web resources referenced from Twitter messages. Abrol et al. developed a system called TWinner, which mines Twitter messages to model users' interests in news topics for improving the quality of Web search [9]. The system analyzes the content of Twitter messages as well as location information in tweets to understand which news topics are popular among the users. Zangerle et al. presented an approach for recommending hashtags to individual users by comparing the tweets where the hashtags occurred and the tweets that a user have published [188]. The

authors used the term frequency and inverse document frequency ($TF \times IDF$) for the comparison of tweets. Chen et al. conducted experiments on recommending URLs posted in Twitter messages and compare strategies for selecting and ranking URLs by exploiting the social network of a user and the general popularity of URLs in Twitter [44]. Similarly, Dong et al. exploited textual features and social network features based on microblogging data to recommend fresh URLs that have possibly not been indexed by Web search engines yet [58]. However, the research efforts mentioned above do not analyze the semantics of the microposts, which is key to understanding the users' microblogging behavior (cf. Chapter 4). In this chapter, we apply our microblogging-based user modeling framework which extracts semantically meaningful topics in the microposts and provides flexible design choices for constructing user profiles in context of different recommender systems. Moreover, by investigating different user modeling strategies in detail, our work provides in-depth insights into user modeling based on microblogging data and its impact on recommender systems.

The real-time nature of information that people published on Twitter poses new challenges for user modeling and personalization. The trending topics as well as individual users' interests evolve over time [111, 115, 138]. Lerman and Ghosh compared the spread of news on Twitter and Digg[1]. The later is a a social news aggregation service that allows users to submit links to news stories and vote on stories submitted by other users [112]. The results suggest that the messages in Twitter are actively spreading over a longer period of time than in Digg. Kwak et al. conducted a temporal analysis of trending topics on Twitter and discovered that over 85% of the tweets posted everyday are related to news [106]. Most Twitter-related research efforts focus on the network structure to analyze the temporal evolution of trending topics or global patterns of of information spread [106, 112, 182]. Our research conducted in this chapter investigates individual microblogging activities and personal interests also in relation to public trends and therefor provides yet unexplored insights.

Given large microblogging datasets, in this chapter we conduct a set of analyses and recommendation experiments to analyze and evaluate the impact of different design dimension and design alternatives on personalization. In particular, we specify the experiments conducted in this chapter as follows.

Analyzing user modeling for news recommendations. We analyze how different topic modeling strategies (see Section 3.2.1) and further semantic enrichment of microblogging activities with external Web resources (e.g. news articles, see Section 4.2) influence the characteristics of user profiles. We also in-

[1] http://www.digg.com

vestigate the temporal patterns that occur in the user profiles. By combining different design dimensions and design alternatives, our user modeling framework provides a variety of user modeling strategies to construct user profiles. In order to evaluate the quality of user profiles, we measure and compare the performance of different user modeling strategies in context of a *personalized news recommender system* (see Section 6.2).

Interweaving trend and user modeling for news recommendations. We demonstrate how to model public trends in microblogging streams and investigate the interplay between personal interests and public trends. Given a large Twitter dataset, we analyze the characteristics of user and trend profiles. We evaluate the trend and user modeling strategies in context of a *personalized news recommender systems* (see Section 6.3).

Analyzing temporal dynamics for URL recommendations We examine how public trends and users's interests in these public trends evolve over time. Moreover, we analyze and evaluate the impact time-sensitive weighting scheme on the performance of a *personalized URL recommender system* (see Section 6.4).

Domain-specific user modeling for product recommendations The user modeling framework allows for generating user profiles for various application settings (cf. Section 3.3). We analyze our user modeling framework in six different application domains. We investigate the quality of user profiles that are adapted to different domains for supporting various recommendation tasks ranging from *product recommendations* to more specific tasks such as *book* or *software product* recommendations (see Section 6.5).

We will answer the following research questions in this chapter.

- How do the different user modeling strategies impact personalized recommendations? Can the temporal patterns be applied to improve recommendation accuracy?
- What is the impact of combining trend and user profiles on the performance of new recommendation systems?
- How are personal interests and user concerns influenced by public trends? How do the time-sensitive weighting functions impact the accuracy of personalized recommender systems?
- To what extent are the domain-specific user modeling strategies beneficial for supporting recommendation systems in different domains?

Table 6.1: Design dimensions and design alternatives that are evaluated in Section 6.2.

design dimension	**design alternatives** (evaluated in Section 6.2)
topic modeling	(i) hashtag-based, (ii) category-based, (iii) entity-based or (iv) LDA-based
enrichment	(i) tweet-only-based enrichment or (ii) linkage and exploitation of external news articles (propagating entities/topics)
temporal constraints	(i) specific time period(s), (ii) temporal patterns (*weekend*, *night*, etc.) or (iii) no constraints

6.2 Analyzing User Modeling on Twitter for Personalized News Recommendation

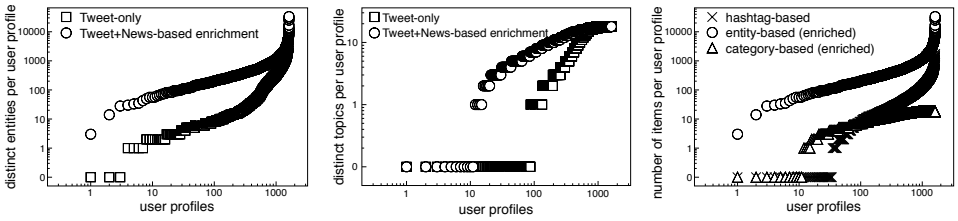
In Chapter 3, we developed a framework that enriches the semantics of individual microblogging activities and provides various strategies for construction user profiles. The characteristics of these user profiles are influenced by different design dimensions and design alternatives. In Chapter 4, we have analyzed the impact of semantic enrichment on the characteristics of constructed user profiles. In this section, we further conduct an in-depth analysis on a large Twitter dataset of more than 2 million tweets to better understand how different topic modeling strategies and temporal constraints impact the characteristics and quality of the resulting user profiles (see Table 6.1). We further measure and compare the impact of the user modeling strategies on the performance of a personalized news recommendation system.

6.2.1 Analysis of Twitter-based User Profiles

To understand how the different user modeling design choices influence the characteristics of the generated user profiles, we applied our Twitter-based user modeling framework to construct user profiles based on a large Twitter dataset. The main research questions to be answered in this analysis can be summarized as follows.

1. How do the different user modeling strategies impact the *characteristics* of Twitter-based user profiles?

2. Which *temporal characteristics* do Twitter-based user profiles feature?



(a) Entity-based profiles (b) Category-based profiles (c) Comparison of different types of profiles

Figure 6.1: Comparison between different user modeling strategies with tweet-only-based or news-based enrichment.

Data Collection and Data Set Characteristics

Over a period of more than two months we crawled Twitter information streams of more than 20,000 users. Together, these people published more than 10 million tweets. To allow for linkage of tweets with news articles we also monitored more than 60 RSS feeds of prominent news media such as BBC, CNN or New York Times and aggregated the content of 77,544 news articles. The number of Twitter messages posted per user follows a power-law distribution. The majority of users published less than 100 messages during our observation period while only a small fraction of users wrote more than 10,000 Twitter messages and one user produced even slightly more than 20,000 tweets (no spam). As we were interested in analyzing also temporal characteristics of the user profiles, we created a sample of 1619 users, who contributed at least 20 tweets in total and at least one tweet in each month of our observation period. This sample dataset contained 2,316,204 tweets in total.

We processed each Twitter message and each news article to identify categories and entities mentioned in the tweets and articles. Further, we applied different linking strategies as proposed in Section 4.2 and connected 458,566 Twitter messages with news articles of which 98,189 relations were explicitly given in the tweets by URLs that pointed to the corresponding news article. The remaining 360,377 relations were obtained by comparing the entities that were mentioned in both news articles and tweets as well as by comparing the timestamps. In Section 4.2 we showed that this method correlates news and tweets with an accuracy of more than 70%. Our hypothesis is that – regardless whether this enrichment method might introduce a certain degree of noise – it impacts the quality of user modeling and personalization positively.

Structural Analysis of Twitter-based Profiles

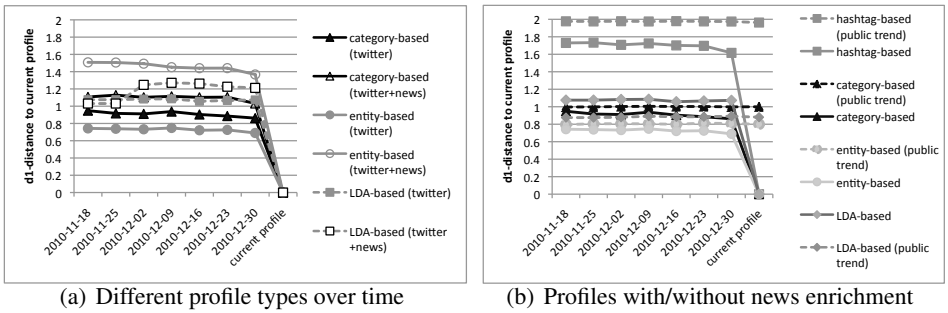
To validate our hypothesis and explore how the exploitation of linked external sources influences the characteristics of the profiles generated by the different user modeling strategies, we analyzed the corresponding profiles of the 1619 users from our sample. In Figure 6.1 we plot the number of distinct (types of) concepts in the topic- and entity-based profiles and show how this number is influenced by the additional news-based enrichment.

For both types of profiles the enrichment with entities and topics obtained from linked news articles results in a higher number of distinct concepts per profile (see Fig. 6.1(a) and 6.1(b)). category-based profiles abstract much stronger from the concrete Twitter activities than entity-based profiles. In our analysis we utilized the OpenCalais taxonomy consisting of 18 topics such as politics, entertainment or culture. The tweet-only-based user modeling strategy, which exploits merely the semantics attached to tweets, fails to create profiles for nearly 100 users (6.2%, category-based) as for these users none of the tweets can be categorized into a topic. By enriching the tweets with topics inferred from the linked news articles we better understand the semantics of Twitter messages and succeed in creating more valuable category-based profiles for 99.4% of the users.

Further, the number of profile facets, i.e. the type of entities (e.g. person, location or event) that occur in the entity-based profiles, increases with the news-based semantic enrichment. While more than 400 twitter-based profiles (more than 25%) feature less than 10 profile facets and often miss entities such as movies or products a user is concerned with, the news-based enrichment detects a greater variety of entity types. For more than 99% of the entity-based profiles enriched via news articles, the number of distinct profile facets is higher than 10.

A comparison of the entity- and category-based user modeling strategies with the hashtag-based strategy (see Fig. 6.1(c)) shows that the variety of entity-based profiles is much higher than the one of hashtag-based profiles. While the entity-based strategy succeeds to create profiles for all users in our dataset, the hashtag-based approach fails for approximately 90 users (5.5%) as the corresponding people neither made use of hashtags nor re-tweeted messages that contain hashtags. Entity-based as well as category-based profiles moreover make the semantics more explicit than hashtag-based profiles. Each entity and topic has a URI which defines the meaning of the entity and topic respectively.

The advantages of well-defined semantics as exposed by the topic- and entity-based profiles also depend on the application context, in which these profiles are used. The results of the quantitative analysis depicted in Fig. 6.1 show that entity- and category-based strategies allow for higher coverage regarding the number of



(a) Different profile types over time (b) Profiles with/without news enrichment

Figure 6.2: Temporal evolution of user profiles: average d_1-distance of current individual user profiles with corresponding profiles in the past.

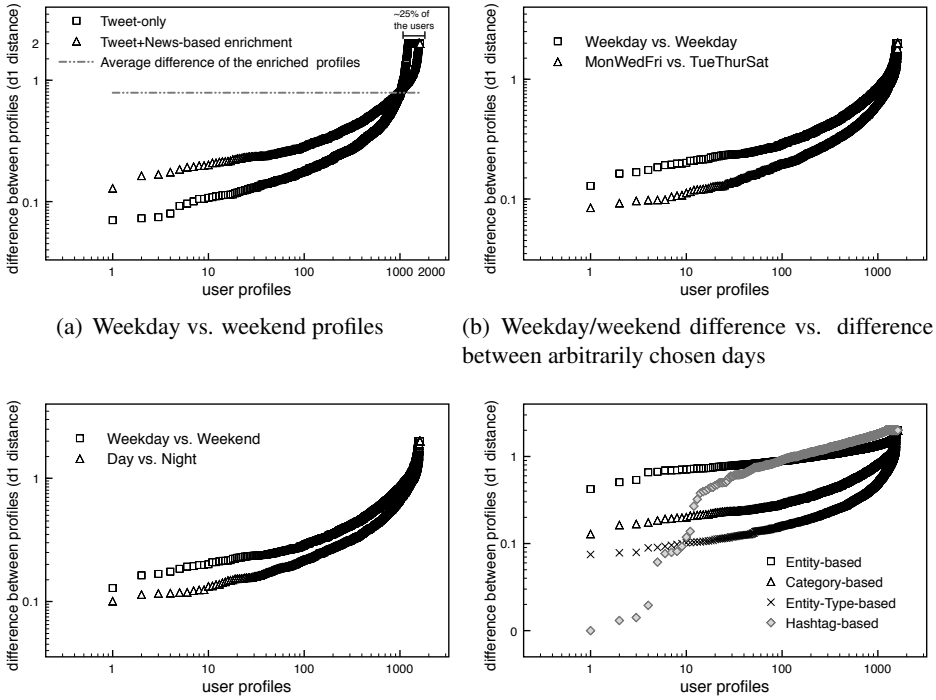
users, for whom profiles can be generated, than the hashtag-based strategy. Further, semantic enrichment by exploiting news articles (implicitly) linked with tweets increases the number of entities and topics available in the profiles significantly and improves the variety of the profiles (the number of profile facets).

Temporal Analysis of Twitter-based Profiles

In the temporal analysis we investigate (1) how the different types of user profiles evolve over time and (2) which temporal patterns occur in the profiles. Regarding temporal patterns we, for example, examine whether profiles generated on the weekends differ from those generated during the week. Similar to the click-behavior analysis by Liu et al. [117], we apply the so-called d_1-*distance* for measuring the difference between profiles in vector representation: $d_1(\vec{p}_x(u), \vec{p}_y(u)) = \sum_i |p_{x,i} - p_{y,i}|$.

The higher $d_1(\vec{p}_x(u), \vec{p}_y(u)) \in [0..2]$ the higher the difference of the two profiles $\vec{p}_x(u)$ and $\vec{p}_y(u)$ and if two profiles are the same then $d_1(\vec{p}_x(u), \vec{p}_y(u)) = 0$. Figure 6.2 depicts the evolution of profiles over time. It shows the average d_1-distance of the current user profiles with the profiles of the same users created based on Twitter activities performed in a certain week in the past. As suggested in [117], we also plotted the distance of the current user-specific profile with the *public trend* (see Fig. 6.2(a)), i.e. the average profile of the corresponding weeks.

For the three different profile types we observe that the d_1-distance slightly decreases over time. For example, the difference of current profiles (first week of January 2011) with the corresponding profiles generated at the beginning of our observation period (in the week around 18th November 2010) is the highest while the distance of current profiles with profiles computed one week before (30th December 2010) is the lowest. It is interesting to see that the distance of the current



(a) Weekday vs. weekend profiles

(b) Weekday/weekend difference vs. difference between arbitrarily chosen days

(c) Weekday/weekend difference vs. day/night difference

(d) Different types of profiles

Figure 6.3: Temporal patterns: comparison between weekend and weekday profiles by means of d_1-distance ((a)-(c): category-based profiles).

profiles with the public trend (i) is present for all types of profiles and (ii) is rather constant over time. This suggests (i) a certain degree of individualism in Twitter and (ii) reveals that the people in our sample follow different trends rather than being influenced by the same trends.

Hashtag-based profiles exhibit the strongest changes over time as the average d_1-distance to the current profile is constantly higher than for the category-, entity-based and LDA-based profiles. Figure 6.2(b) discloses that entity-based profiles change stronger over time than category-based profiles when news-based enrichment is enabled. When merely analyzing Twitter messages one would come to a different (possibly wrong) conclusion (see Fig. 6.2(a)).

Figure 6.3 illustrates temporal patterns we detected when analyzing the individual user profiles. In particular, we investigate how profiles created on the weekends differ from profiles (of the same user) created during the week. For category-based profiles generated solely based on Twitter messages, it seems that for some users

the weekend and weekday profiles differ just slightly while for 24.9% of the users the d_1-distance of the weekend and weekday profile is maximal (2 is the maximum possible value, see Fig. 6.3(a)). The news-based enrichment reveals however that the difference of weekend and weekday profiles is a rather common phenomenon: the curve draws nearer to the average difference (see dotted line); there are less extrema, i.e. users for whom the d_1-difference is either very low or very high. Hence, it rather seems that the tweets alone are not sufficient to get a clear understanding of the users concerns and interests.

Fig. 6.3(b) further supports the hypothesis that weekend profiles differ significantly from weekday profiles. We observer that the corresponding distances $d_1(\vec{p}_{weekend}(u), \vec{p}_{weekday}(u))$ are consistently higher than the differences of profiles generated on arbitrarily chosen days during the week. This *weekend pattern* is more significant than differences between category-based profiles generated based on Twitter messages that are either posted during the evening (6pm-3am) or during the day (9am-5pm) as shown in Fig. 6.3(c). Hence, the individual topic drift – i.e. change of topics individual users are concerned with – between day and evening/night seems to be smaller than between weekdays and weekends.

The weekend pattern is coherent over the different types of profiles. Different profile types however imply different drift of interests or concerns between weekend and weekdays (see Fig. 6.3(d)). Hashtag-based and entity-based profiles change most while the types of entities people refer to (persons, products, etc.) do not differ that strongly. When zooming into the individual entity-based profiles we see that entities related to leisure time and entertainment become more important on the weekends.

The temporal analysis thus revealed two important observations. First, user profiles change over time: the older a profile the more it differs from the current profile of the user. The actual profile distance varies between the different types of profiles. Second, weekend profiles differ significantly from weekday profiles.

6.2.2 Exploitation of User Profiles for Personalized News Recommendations

We further investigate the impact of the different user modeling strategies on recommending news articles:

1. To which degree are the profiles created by the different user modeling strategies appropriate for recommending news?
2. Can the identified (temporal) patterns be applied to improve recommendation accuracy?

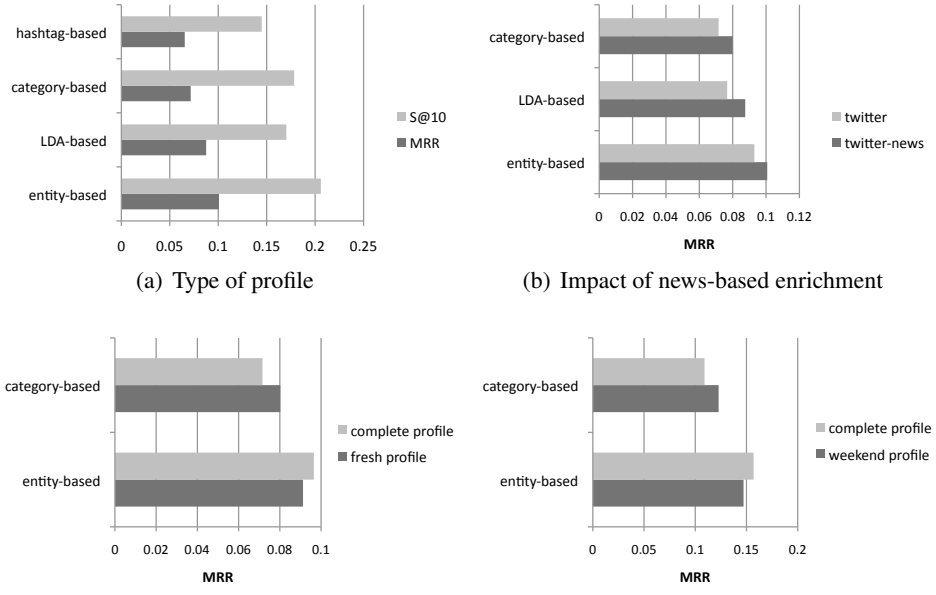
News Recommender System and Evaluation Methodology

Recommending news articles is a non-trivial task as the news items, which are going to be recommended, are *new* by its very nature, which makes it difficult to apply collaborative filtering methods, but rather calls for content-based or hybrid approaches [117]. Our main goal is to analyze and compare the applicability of the different user modeling strategies in the context of news recommendations. We do not aim to optimize recommendation quality, but are interested in comparing the quality achieved by the same recommendation algorithm when inputting different types of user profiles. Therefore we apply a lightweight content-based algorithm that recommends items according to their cosine similarity with a given user profile. We thus cast the recommendation problem into a search and ranking problem where the given user profile, which is constructed by a specific user modeling strategy, is interpreted as query.

Definition 10 (Recommendation Algorithm) *Given a user profile vector $\vec{p}(u)$ and a set of candidate news items $N = \{\vec{p}(n_1), ..., \vec{p}(n_n)\}$, which are represented via profiles using the same vector representation, the recommendation algorithm ranks the candidate items according to their cosine similarity to $\vec{p}(u)$.*

$$sim_{cosine}(\vec{p}(u), \vec{p}(n_i)) = \frac{\vec{p}(u) \cdot \vec{p}(n_i)}{||\vec{p}(u)|| \cdot ||\vec{p}(n_i)||} \tag{6.1}$$

Given the Twitter and news media dataset described in Section 6.2.1, we considered the last week of our observation period as the time frame for computing recommendations. The ground truth of news articles, which we consider as *relevant* for a specific user u, is obtained via the Twitter messages (including re-tweets) posted by u in this week that explicitly link to a news article published by BBC, CNN or New York Times. We thereby identified, on average, 5.5 relevant news articles for each of the 1619 users from our sample. For less than 10% of the users we found more than 20 relevant articles. The candidate set of news articles, which were published within the recommendation time frame, contained 5529 items. We then applied the different user modeling strategies together with the above algorithm (see Def. 10) and set of candidate items to compute news recommendations for each user. The user modeling strategies were only allowed to exploit tweets published before the recommendation period. The quality of the recommendations was measured by means of *MRR* (Mean Reciprocal Rank), which indicates at which rank the first item relevant to the user occurs on average, and $S@k$ (Success at rank k), which stands for the mean probability that a relevant item occurs within the top k of the ranking. In particular, we will focus on S@10 as our recommendation system will list 10 recommended news articles to a user. We tested statistical significance of



(a) Type of profile

(b) Impact of news-based enrichment

(c) Fresh vs. complete profile history

(d) Weekend recommendations

Figure 6.4: Results of news recommendation experiment.

our results with a two-tailed t-Test where the significance level was set to $\alpha = 0.01$ unless otherwise noted.

Results

The results of the news recommendation experiment are summarized in Fig. 6.4 and validate findings of our analysis presented in Section 6.2.1. Entity-based user modeling (with news-based enrichment), which produces according to the quantitative analysis (see Fig. 6.1) the most valuable profiles, allowed for the best recommendation quality and performed significantly better than hashtag-based user modeling (see Fig. 6.4(a)). category-based user modeling also performed better than the hashtag-based strategy – regarding S@10 the performance difference is significant. Since the category-based strategy models user interests within a space of 18 different topics (e.g., politics or sports), it further required much less run-time and memory for computing user profiles and recommendations than the hashtag- and entity-based strategies, for which we limited dimensions to the 10,0000 most prominent hashtags and entities respectively. The LDA-based user modeling strategy slightly performed slightly better hashtag- and category-based strategies with respect to MRR. However, it also requires more resource (e.g, run-time and mem-

ory) to train the topic model for constructing the LDA-based user profiles.

Further enrichment of category- LDA-, and entity-based profiles with categories and entities extracted from linked news articles, which results in profiles that feature more facets and information about users' concerns (cf. Section 6.2.1), also results in a higher recommendation quality (see Fig. 6.4(b)). Exploiting both tweets and linked news articles for creating user profiles improves MRR significantly ($p < 0.05$). In Section 6.2.1 we observed that user profiles change over time and that recent profile information approximates future profiles slightly better than old profile information. We thus compared strategies that exploited just recent Twitter activities (two weeks before the recommendation period) with the strategies that exploit the entire user history (see Fig. 6.4(c)). For the category-based strategy we see that *fresh* user profiles are more applicable for recommending news articles than profiles that were built based on the entire user history. However, entity-based user modeling enables better recommendation quality when the complete user history is applied. Results of additional experiments [3] suggest that this is due to the number of distinct entities that occur in entity-based profiles (cf. Fig. 6.1): long-term profiles seem to refine preferences regarding entities (e.g. persons or events) better than short-term profiles.

In Section 6.2.1 we further observed the so-called *weekend pattern*, i.e. user profiles created based on Twitter messages published on the weekends significantly differ from profiles created during the week. To examine the impact of this pattern on the accuracy of the recommendations we focused on recommending news articles during the weekend and compared the performance of user profiles created just by exploiting weekend activities with profiles created based on the complete set of Twitter activities (see Fig. 6.4(d)). Similarly to Fig. 6.4(c) we see again that the entity-based strategy performs better when exploiting the entire user history while the category-based strategy benefits from considering the weekend pattern. For the category-based strategy recommendation quality with respect to MRR improves significantly when profiles from the weekend are applied to make recommendations during the weekend.

6.2.3 Synopsis

We conducted a large scale analysis on a Twitter dataset to investigated how the different design alternatives influence the characteristics of the generated user profiles. Given a large dataset consisting of more than 2 million tweets we created user profiles and revealed several advantages of semantic entity- and category-based topic modeling strategies, which exploit the full functionality of our Twitter-based user modeling framework, over hashtag-based topic modeling strategies. We saw

that further enrichment with semantics extracted from news articles, which we correlated with the users' Twitter activities, enhanced the variety of the constructed profiles and improved accuracy of news article recommendations significantly.

Further, we analyzed the temporal characteristics of the different types of profiles. We observed how profiles change over time and discovered temporal patterns such as characteristic differences between weekend and weekday profiles. We also showed that the consideration of such temporal characteristics is beneficial to recommending news articles when dealing with category-based profiles while for entity-based profiles we achieve better performance when incorporating the entire user history.

6.3 Interweaving Trend and User Modeling on Twitter for Personalized News Recommendation

In previous section, we analyzed the impact of different design dimensions and design alternatives on the characteristics of user profiles constructed based on Twitter activities. We investigated how different user modeling strategies impact personalization and discovered that the consideration of temporal profile patterns can improve recommendation quality. In this section, we research whether the microblogging activities can be exploited to generate profiles that reflect the interests of a user in current trending news topics. We present an approach to model trends on Twitter and evaluate different strategies for generating trend-aware user profiles which are used in personalized news recommendations. Therefore, we:

- introduce strategies for identifying trends on Twitter and describe trend modeling that allows for the generation of semantic trend profiles by exploiting temporal dynamics of Twitter activities,
- demonstrate how both trend and user profiles can be combined to model users' personal interests in relation to current Twitter trends,
- analyze the temporal dynamics of user profiles and trend profiles and
- evaluate trend and user modeling strategies for recommending news articles to users and prove the effectiveness of our strategies.

6.3.1 Trend Modeling on Twitter

While user profiles represent personal interests of a specific user, trend profiles describe the trending interests of the entire user community. In line with the different

topic modeling strategies for user profile construction, trends may also refer to hashtags, category or entities. In the following, we describe how one can identify and model trends on Twitter within a given period of time. A given time interval j, for which trends should be extracted and modeled, is defined as a tuple that consist of a start and end timestamp:

$$I_j =< ts_{start}, ts_{end} > \tag{6.2}$$

Here, ts_{start} and ts_{end} denote the timestamps that specify the beginning and end of the j-th time interval respectively. Hence, the generic model of profiles that represent public trends for a given time interval can be defined analogously to the user profile model (cf. Definition 1):

Definition 11 (Trend Profile) *The* trend profile $T(I_j)$ *for a given time interval I_j is a set of weighted concepts where for a concept $c \in C$ its weight $w(I_j,c)$ is computed by a certain function w.*

$$T(I_j) = \{(c, w(I_j,c)) | c \in C\}$$

Here, C denotes the set of candidate concepts from which the trends can be extracted in the given time interval I_j. With $\vec{t}(I_j)$ we refer to $T(I_j)$ in its vector space model representation, where the value of the i-th element refers to $w(I_j,c_i)$ and the values of all elements are normalized to make the sum of values to 1.

Our approach for modeling trends allows for three types of concepts that imply different types of trend profiles: hashtag-based, category-based and entity-based profiles. The weighting function $w(I_j,c)$ is applied to measure the importance and popularity of a concept in a specific period of time. In particular, we make use of term frequency and inverse document frequency and also introduce time-sensitive variations of these measures. Before, we introduce the time-sensitive weighting schemes we first describe how we exploit the term frequency based methods to deduce the weight for a given concept in a certain time interval.

- **TF:** For a given time interval I_j, the term frequency TF of a concept c is the fraction of concept references that refer to c.

$$w_{TF}(I_j,c) = \frac{n_{c,j}}{\sum_{c \in C} n_{c,j}} \tag{6.3}$$

  where $n_{c,j}$ denotes the number of (enriched) tweets that refer to concept c during time interval I_j.

- **TF×IDF:** The inverse document frequency (IDF) can be applied to value the specificity of a concept c within a given period of time I_j.

$$w_{TF\times IDF}(I_j,c) = w_{TF}(I_j,c)\cdot log(\frac{|I|}{1+|\{I_i:n_{c,i}>0\}|}) \tag{6.4}$$

  where $w_{TF}(I_j,c)$ is the term frequency of concept c in the time interval I_j, $|I|$ denotes the number of separated time intervals and $|\{I_i : n_{c,i} > 0\}|$ is the number of time intervals in which the concept c was referenced at least once.

Our analysis of the Twitter activities (see Section 6.3.2) reveals that there are concepts (entities) such as "United States" that show constant popularity and are not specific for certain time periods. In contrast, there exist other new occurring entities that quickly spread among Twitter users and gain popularity in a short period of time and fade out after some time. For example, "WikiLeaks" was popular for around two weeks and then gradually decreased in popularity. The $TF \times IDF$ weighting function values such specificity of concepts with respect to a set of time intervals I. $TF \times IDF$ is therefore a discrete measure that heavily depends on the definition of time intervals – in our evaluation we experimented with time intervals of one week.

To measure the temporal dynamics of a concept c on a more continuous spectrum, we calculate the standard deviation of the timestamps of (semantically enriched) tweets that refer to c [93]. Hence, we apply the corrected sample standard deviation as follows:

$$\sigma(c) = \sqrt{\frac{\sum_{k=1}^{N}(ts_k - \overline{ts})^2}{N-1}} \tag{6.5}$$

Here, ts_k is the timestamp of the k-th tweet that refers to concept c, $\overline{ts}$ is the average timestamp of tweets that relate to c and N is the overall number of tweets that refer to c. Huang et al. use a similar interpretation to characterize the temporal stability of hashtags [93]. Based on the above interpretation of standard deviation, one can expect the following behavior: the lower the value of $\sigma(c)$ the shorter is the time period in which concept c is referenced by tweets. If a concept c is just mentioned once then the standard deviation will be zero ($\sigma(c) = 0$). In contrast, the higher the standard deviation, the more constantly is a concept referenced from tweets. Given this notion of standard deviation, we modify the conventional TF and $TF \times IDF$ weighting functions and introduce two new time-sensitive weighting schemes.

- **Time-sensitive TF:** For a given time interval I_j, the time-sensitive term fre-

quency TF of a concept c is defined as follows:

$$w_{t-TF}(I_j,c) = w_{TF}(I_j,c) \cdot (1 - \hat{\sigma}(c)) \tag{6.6}$$

where $w_{TF}(I_j,c)$ is the term frequency of concept c for time interval I_j and $\hat{\sigma}(c)$ denotes the normalized standard deviation $\sigma(c)$ which is calculated via:

$$\hat{\sigma}(c) = \frac{\sigma(c)}{\max\{\sigma(c_i) : c_i \in C\}} \tag{6.7}$$

- **Time-sensitive TF×IDF:** Similarly, the time-sensitive $TF \times IDF$ is specified as follows:

$$w_{t-TF\times IDF}(I_j,c) = w_{TF\times IDF}(I_j,c) \cdot (1 - \hat{\sigma}(c)) \tag{6.8}$$

where $w_{TF\times IDF}(I_j,c)$ denotes the weight fro conventional $TF \times IDF$ function and $\hat{\sigma}(c)$ denotes normalized standard deviation value for concept c.

Hence, $\hat{\sigma}(c) \in [0..1]$ and in particular the factor $(1-\hat{\sigma}(c))$ is used to de-emphasize TF and $TF\times IDF$. The higher the normalized standard deviation $\hat{\sigma}(c)$, the lower the weight $w_{t-TF}(I_j,c)$ and $w_{t-TF\times IDF}(I_j,c)$ respectively.

We use $T_{w@k}(I_j)$ to denote the trend profile that selects the top k weighted concepts $c \in C$ when applying a certain weighting function w. For example, the trend profile $T_{t-TF@100}(I_j)$ contains the top 100 trending concepts in C based on the weights of concepts that are calculate by time-sensitive TF weighting function.

Combining Trends and Personal User Profiles

Based on the framework components discussed above, we can generate two kinds of profiles: (i) user profiles which represent personal interests and (ii) trend profiles which represent the public trends in Twitter for a specific period of time. Our main goal is to investigate how to better reflect personal interests in current news topics for supporting personalization. In particular, we are interested in systems that aim for trend-aware personalization like news recommendation systems that deal with new items and aim for recommending these new items to users. For this purpose, we propose to interweave trend and user profiles to model users' interests in the context of current Twitter trends (see Figure 6.5).

Both trend and user profiles benefit from the linkage and semantic enrichment discussed above and allow for the generation of hashtag-based, category-based and entity-based profiles. While the user modeling component expects the history of

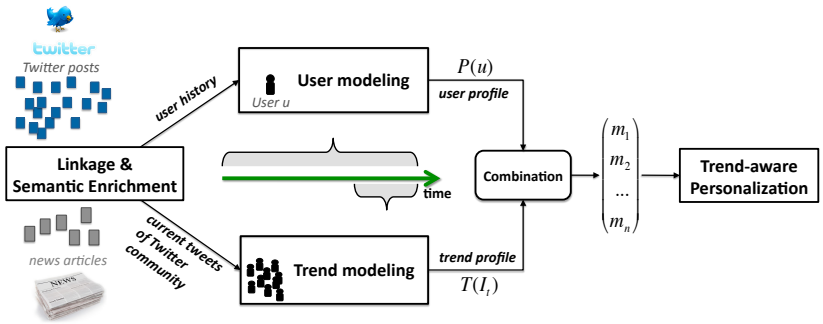


Figure 6.5: Combining trend and user modeling to support trend-aware personalization

Twitter activities of a specific user as input, the trend identification and trend modeling is based on the tweets published by the entire Twitter community in a certain period in time (see Figure 6.5). By combining the personal user profiles with the trend profiles, we aim to generate a profile that better estimates user interests in relation to public trends for supporting trend-aware personalization. We therefore apply a classical mixture method to combine a given user profile with a trend profile of a certain time period I_j.

$$\vec{m}(I_j,u) = d * \vec{p}(u) + (1-d) * \vec{t}(I_j) \tag{6.9}$$

Here, $\vec{p}(u)$ and $\vec{t}(I_j)$ denote the vector representation of a user profile $P(u)$ and trend profile $T(I_j)$ respectively. With $\vec{m}(I_j,u)$ we refer to the combined profile $M(I_j,u)$ in its vector space model representation. The parameter $d \in [0..1]$ allows for adjusting the influence of the user profile and trend profile on the combined profile. By increasing d, we can emphasize the user profile and de-emphasize the trend profile at the same time. In the following sections, we focus on entity-based profiles and analyze how different configurations for combining trend and user profiles impact the quality of trend-aware news recommendation system.

6.3.2 Temporal Analysis of User and Trend Profiles on Twitter

Based on the large Twitter dataset used in Section 6.2, we analyze the temporal dynamics of user and trend profiles to answer the following research questions:

1. How do interests change for different types of entities?
2. How do trends evolve over time and how does the trend modeling strategies impact the characteristics of trend profiles?

Temporal Analysis of User Profiles

In Section 6.2.1, we applied d_1-distance to analyze how user profiles evolve over time. Our results reveal that entity-based profiles change stronger over time than other types of user profiles (see Figure 6.2). New entities may occur over time and the entity-based profiles allow for capturing interests into these new entities. Hence, entity-based profiles seem to capture current interests of a user more precisely than hashtag-, category-, or LDA-based profiles.

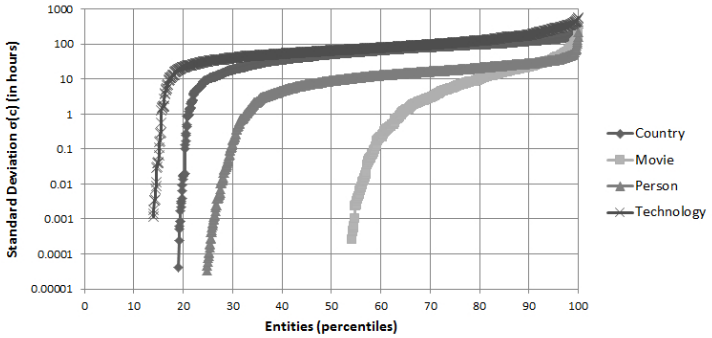


Figure 6.6: Standard deviation of different types of entities (cf. Equation 6.11).

To answer the first research question, we plotted in Figure 6.6 the standard deviation (see Equation 6.11) of different types of entities. We observe that different types of entities have a different life span within entity-based profiles. The higher the standard deviation of a certain entity is, the more consistently does the entity occur in Twitter messages posted by the users. For example, we see that entities of type country or technology have, on average, a higher standard deviation than movies or persons, i.e. country or technology entities occur more constantly within profiles than movie and person entity. In fact, for more than 50% of the movies standard deviation is 0, which means that the corresponding entities are mentioned just once by a user. In summary, we observe that some fractions of the entity-based profiles are rather constant while others change dynamically over time.

Temporal Analysis of Trend Profiles

We now turn to the second question to investigate the temporal characteristics of trends in Twitter and analyze the effectiveness of trend modeling strategies. In Figure 6.7 we plot the occurrence frequency of popular entities over time. Some entities such as “United States” or “USD” are continuously among the most frequently mentioned entities. We assume that interest in those entities will anyhow be captured by the long-term user profiles. Hence, when generating trend profiles, we are rather in-

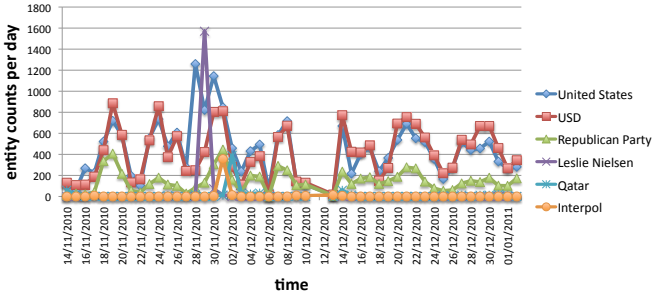


Figure 6.7: Top entities for a given week.

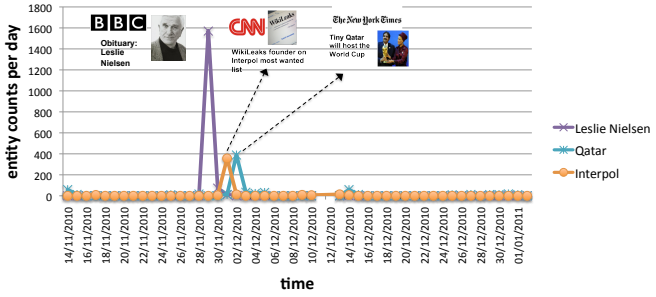


Figure 6.8: Top trends for a given week.

terested in trending entities that have a peak at a certain point of time. For example, Figure 6.7 shows a peak for "Leslie Nielson" who was usually not mentioned frequently but suddenly became the most popular entity at the end of November 2010. While this trend is clearly visible, other trends are overloaded by entities that are constantly popular. Using the time-sensitive weighting function (cf. Section 6.3.1) in particular we are able to filter out those popular entities and can identify trending entities which are of particular importance for a specific period in time.

Figure 6.8 shows the entities from Figure 6.7 which are identified as trending entities according to the time-sensitive $TF \times IDF$ weighting function. We observe that there are further peaks for "Qatar" and "Interpol" at the beginning of December 2010 which are overloaded by "United States" and "USD" in Figure 6.7. While the death of Leslie Nielson became a trending topic on Twitter, the news that Qatar will host the World Cup or that Interpol declared William Assange, the founder of WikiLeaks, as one of the most wanted persons were rather trending side topics. By applying the time-sensitive weighting function, we can thus discover such trending entities and generate profiles that capture trends a user is interested in.

6.3.3 Evaluation of Trend and User Modeling for Recommending News Articles

In this experiment, we investigate the impact of the trend and user modeling strategies on the quality of personalized news recommender systems and answer the following research questions.

1. Which weighting function is best for generating trend profiles in the context of recommending news articles?
2. What kind of profiles – personal user profiles or public trend profiles – are best for recommending news articles?
3. Can we improve recommendation performance by combining trend and user profiles?

News Recommender and Evaluation Methodology

The user and trend profiles are represented in a semantically meaningful way so that they can be consumed by applications that require personal user interests as well as current trends. Here, we evaluate our trend and user modeling strategies for supporting personalized news recommendations.

Given user profiles which represent the long-term personal interests, trend profiles which represent rather short-term public trends and combined profiles, our goal is to analyze the impact of these profile modeling methods on the news recommendation quality. We apply the content-based algorithm that recommends items according to their cosine similarity with a given profile generated by the trend and user modeling framework.

Definition 12 (Recommendation Algorithm) *Given a profile $\vec{p}$ in vector representation, which either represents a user profile $P(u)$, a trend profile $T(I_j)$ or a combined profile $M(I_j,u)$, and a set of candidate news items $N = \{\vec{p}_{news}(n_1),...,\vec{p}_{news}(n_n)\}$, which are represented via profiles using the same vector representation as $\vec{p}$, the recommendation algorithm ranks the candidate items according to their cosine similarity to $\vec{p}$:*

$$sim_{cosine}(\vec{p},\vec{p}_{news}(n_i)) = \frac{\vec{p}\cdot\vec{p}_{news}(n_i)}{||\vec{p}||\cdot||\vec{p}_{news}(n_i)||} \tag{6.10}$$

Given the dataset described in Section 6.2, we considered the last week of our observation period as the time interval, denoted as I_r for computing recommendations. The ground truth of news articles, which we consider as *relevant* for a specific

user u, is obtained via the Twitter messages (including re-tweets) posted by u in this week that explicitly link to a news article published by BBC, CNN or New York Times. We run our experiments for these 577 users, for whom we identified at least five relevant news articles during our recommendation period. For each of these users, we compare the three alternative modeling strategies for generating an input profile to be fed into the news recommendation algorithm: the user profile $P(u)$, the trend profile $T(I_r)$, and the combined profile $M(I_r,u)$ (see Section 6.3.1). $P(u)$ is generated based on the enriched tweets that were published by u before the start of the recommendation period. $T(I_r)$ is generated for the recommendation period I_r and we compare the different weighting functions as described in Section 6.3.1 to identify the best trend modeling strategy. For the combined profile $M(I_r,u)$, we further experiment with different configurations for selecting the top trending concepts and vary the parameter d when combining personal profiles and trend profiles to investigate the influence of the different types of profiles on the news recommendation task. The quality of the recommendations is measured by means of *MRR* (Mean Reciprocal Rank), which indicates at which rank the first item relevant to the user occurs on average, and *S@k* (Success at rank k), which stands for the mean probability that a relevant item occurs within the top k of the ranking. We tested statistical significance of our results with a two-tailed t-Test where the significance level was set to $\alpha = 0.01$ unless otherwise noted.

Results

The results of the news recommendation experiments are demonstrated in Figure 6.9, 6.10 and 6.11. Each figure answers one of the research questions raised at the beginning of this section.

First, we investigate which weighting function is best for generating trend profiles in the context of news recommendations. In Section 6.3.1, we proposed four types of weighting functions for the generation of trend profiles. Our hypothesis is that our time-sensitive methods that adjust conventional TF and $TF \times IDF$ weighting functions by means of standard deviation allow us to better emphasize the emerging and popular concepts in a specific period of time. Hence, we assume that the time-sensitive weighting schemes allow us to achieve higher recommendation quality. To validate our hypothesis, we generate trend profiles by applying different weighting functions and compare *MRR* and *S@5* measures for the recommendations. Figure 6.9 reveals that the time-sensitive weighting functions improve the quality of news recommendations clearly. With the time-sensitive $TF \times IDF$ weighting function we reach the best recommendation performance and improve over the TF baseline by 9.3% and 40.1% with respect to MRR and S@5 respectively. These results thus confirm our hypothesis and show that time-sensitive weighting func-

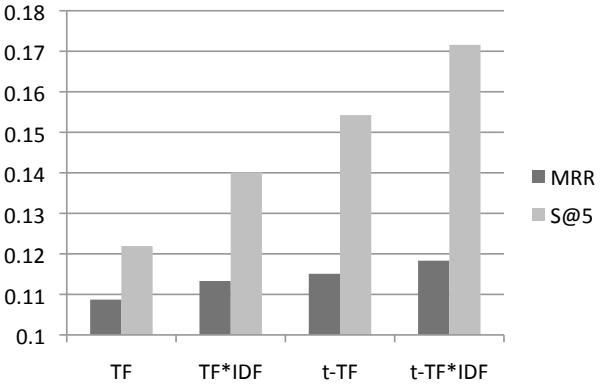


Figure 6.9: Comparison of different weighting functions (for trend profile $T_{w@500}(I_r)$)

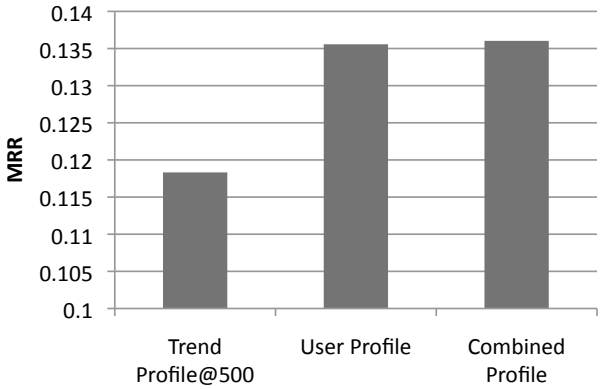


Figure 6.10: Comparison of different type of profiles ($T_{t-TF\times IDF@500}(I_r)$, $P(u)$ and $M(I_r,u)$)

tions and time-sensitive $TF\times IDF$ in particular are best for generating trend profiles in the context of recommending news articles which answer the first research question raised at the beginning of this section.

Second, to answer what kind of profiles are best for news recommendations, we compare the performance of the three types of profiles that are generated by our trend and user modeling framework: user profiles $P(u)$, trend profiles $T(I_r)$ and the mixture $M(I_r,u)$ of both of these profiles. By interweaving the personal interests into global trends that we detect in the Twitter community, we expect to better estimate the current and future interests for supporting trend-aware personalization and news article recommendations in particular. Figure 6.10 compares the performance of the strategies with respect to MRR when using $TF\times IDF$ as weighting function

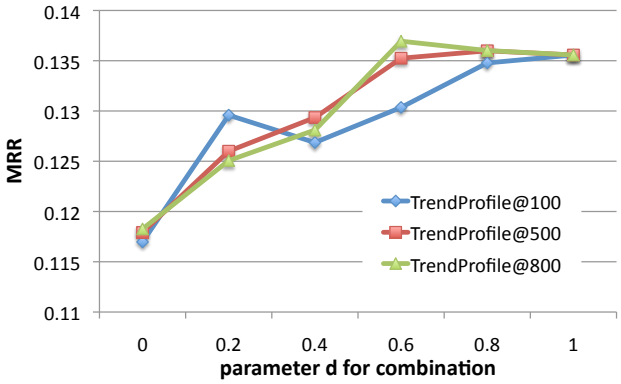


Figure 6.11: Comparison of different strategies for combining user and trend profiles

for modeling the top 500 trends ($T_{t-TF\times IDF@500}(I_r)$). We observe that user profiles allow for better recommendation quality. In fact, personal user interest profiles $P(u)$ improve the performance of public trend profiles $T_{t-TF\times IDF@500}(I_r)$ by 15.3% regarding mean reciprocal rank of the first relevant item ($p < 0.05$). Furthermore, by combining personal interests with public trends, the recommendation quality can be improved slightly.

Finally, to further investigate how the combination of user and trend profiles impacts personalized news recommendations, we evaluate the mixture strategy of combining trend and user profiles for different configurations. Figure 6.11 shows the results for varying parameter d when combining $P(u)$ and $T(I_r)$ (cf. Equation 6.9) and moreover for a varying the number of top k concepts selected for generating the trend profiles. The results reveal that personal user interest profiles seem to be more important than public trends as the recommendation quality with respect to MRR increases when the influence of $P(u)$ is increased. When combining $P(u)$ with $T_{t-TF\times IDF@800}(I_r)$ we achieve a global maximum in performance for $d = 0.6$ which also clearly improves over the strategy that is merely based on personal user interests ($d = 1$, $P(u)$). We conclude that user profiles are more important for generating personalized news recommendations. However, by combining trend and user profiles we achieve the best recommendation performance.

6.3.4 Synopsis

We presented an approach for integrating trend and user modeling on Twitter. Our approach features functionality for enriching the semantics of tweets and therefore allows for the generation of semantically meaningful profiles which represent both

personal interests and public trends. Those profiles can be re-used outside of Twitter by other applications that aim for trend-aware personalization.

We evaluated the trend and user modeling strategies in the context of news article recommendations. Based on a large dataset of more than 10 million tweets, we have analyzed trend and user profiles and saw that user profiles dynamically change over time. In particular, interests in persons or movies vary stronger over time than interests in locations. Furthermore, we proposed time-sensitive strategies allow for the discovery of local trends that are of particular importance in a specific period of time and we saw that our time-sensitive variation of $TF \times IDF$ achieves the best trend modeling performance in the context of news recommendations. Given the trend and user modeling strategies, we showed that personal user interest profiles are more important for the news article recommendation process than public trends. By interweaving trend and user profiles we succeeded in further improving the recommendation quality.

6.4 Analyzing Temporal Dynamic on Twitter for Personalization

Our user modeling framework also features different weighting schemes ranging from term frequency based methods that count the number of occurrences of topics to more advanced time-sensitive weighting function schemes that incorporate a temporal decay for assigning a weight to a topic of interest (cf. Section 3.2.4). In this section, we investigate the following research questions that concern the temporal evolution of individual user profiles inferred from Twitter activities and the impact of time-sensitive weighting schemes on the quality of user profiles in context of a personalized URL recommender system.

- To what extend do personal interests vary over time?
- How are personal interests and user concerns influenced by public trends? Do inferred interests profiles allow for predicting which trends will be adopted by a user?
- How does the time-sensitive weighting function impact the accuracy of personalization?

6.4.1 Evolution of User Interests in Trending Topics

To better illustrate how the interests of users into a topic discussed on Twitter change over time, we start with a concrete example trending topic, the Egyptian revolution[2], which started on January 25th, 2011. In this analysis, we aim to study (i) how trending topics evolve over time and (ii) how the interest of individual users into a topic change over time.

For representing a trending topic, it is often not sufficient to represent it just via a single concept such as a hashtag (words starting with "#"). For example, regarding the Egyptian revolution a hashtag like "#egypt" could be considered as a representative concept to describe this topic. However, (i) not all tweets that contain the hashtag "egypt" refer to the revolution in Egypt and (ii) there exist tweets that refer to the revolution but do not mention the hashtag "#egypt". Instead, other terms that refer to entities such as *Mubarak* (person) or *Cairo* (location) may be used. Therefore, we use the model described in Definition 11 (Section 6.3) to represent a trending topic on Twitter as a set of weighted concepts where a concept may refer to an arbitrary entity and where the weight indicates how important the concept is for the topic.

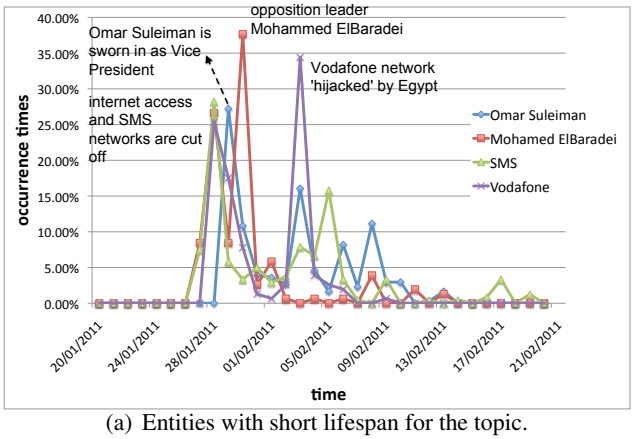
The above model for creating the representation of a topic expects a timestamp as input because concepts that relate to a certain topic may change over time. On the one hand, the importance of concepts could vary at different points in time and, on the other hand, new concepts could arise while other concepts that were once representing the topic could entirely become useless to describe the topic. Hence, the representation of a topic depends on the time when the profile is demanded.

Evolution of Topics over Time

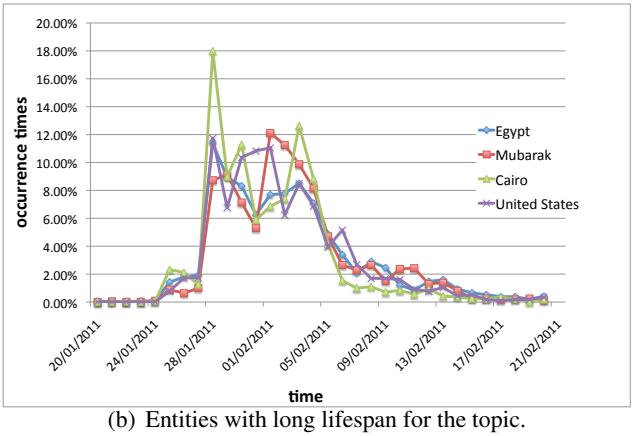
To analyze how the topic *Egyptian revolution* evolves over time, we selected popular entities on every day of our observation period based on their co-occurrence frequency with hashtags such as "#jan25" or "#tahrir" which we could almost unambiguously relate to the topic.

Figure 6.12 illustrates how the occurrence frequency of entities, which are related to the Egyptian revolution, changes over time. Some entities like *Cairo* and *Mubarak* are popular for this topic over a long period in time (see Figure 6.12(b)), which means that Twitter users continuously refer to these entities when publishing tweets about the topic. The occurrence frequencies of these entities quickly reach their peaks three days after the beginning of the Egyptian revolution, which started

[2] http://en.wikipedia.org/wiki/Timeline_of_the_2011_Egyptian_revolution



(a) Entities with short lifespan for the topic.



(b) Entities with long lifespan for the topic.

Figure 6.12: Relative occurrence frequencies of entities related to the Egyptian revolution.

on January 25th, and then decrease rather slowly over the next two weeks.

In contrast, the entities shown in Figure 6.12(a) show burst-like spikes and seem to be relevant for the topic only for a short period in time. For example, many messages that were posted on January 28th were related to the entity *SMS* and referred to the shutdown of the Internet access and short messaging services in Egypt that happened on January 26th, such as the following tweet:

> "Again, latest Egypt updates: internet shut down, SMS and Blackberry down, plainclothes police setting cars on fire"

Therefore the entity *SMS* became very popular on that day. Similarly, *Omar Suleiman* was mentioned in many messages on January 30th as he was sworn in as vice pres-

ident on January 29th which resulted in further protests as reported in the following message:

> "Al Jazeera breaking: Protesters loudly condemn the appointment of Omar Suleiman as Vice President"

Similarly, the leader of the opposition *Mohamed ElBaradei* became popular for the topic on January 30th as well. Moreover, the peak for *Vodafone* on February 3rd is much likely related to the news that the mobile phone company announced that the Egyptian authorities had hijacked Vodafone's network.

Our analysis presented in Figure 6.12 thus demonstrates that the importance of entities for a given topic changes over time. While there are some entities that are continuously good representatives for a topic (e.g. *Mubarak*), there are other entities (e.g. *SMS*) which characterize a topic only for a short period in time. When creating the representation of a topic it is thus reasonable to consider multiple concepts (e.g. entities and hashtags) and to compute the importance of each concept as a function of the time when the topic representation is requested.

Evolution of User Interests into Topics

Having seen how a topic is discussed within a community of users and how the representation of a topic emerges and changes over time, we now analyze how the interests of individual users into a topic evolves over time. We therefore selected a subset of 1619 Twitter users. In particular, those users for which we monitored at least 20 Twitter messages in total and observed at least 10 Twitter messages during the time of the Egyptian revolution but not necessarily 10 messages that are related to the incident in Egypt. In fact, we discovered that 70% of the sample users showed interest into the Egyptian revolution, i.e. 70% of the users (re-)tweeted a message that was mentioning a concept of the corresponding topic representation. While these users were interested in the topic, the individual behavior showed interesting specifics. For example, not all the users started tweeting about the event from the very beginning (January 25th). Figure 6.13 shows for each day the number of users who published their first tweet about the Egyptian revolution and therefore showed for the first time that they are – to some extent – concerned with the topic.

As shown in Figure 6.13, most people do not join the discussion or dissemination of the event immediately after it happens. While the small amplitudes before January 25th can be considered as noise and seem to be caused by the modeling of the topic, on the day of the first wave of protest in Egypt, the "Day of Revolt", slightly less than 150 of our sample users joined the discussion on the topic. After

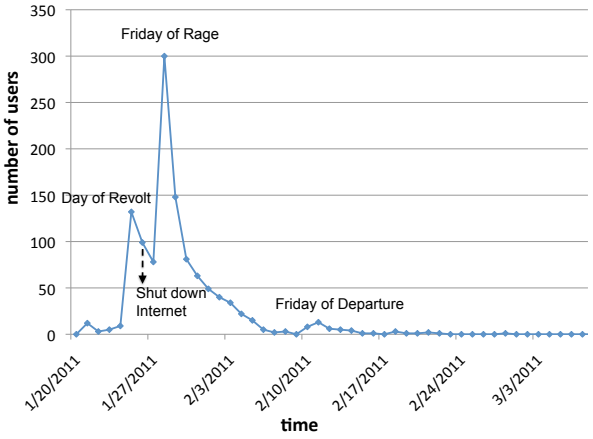
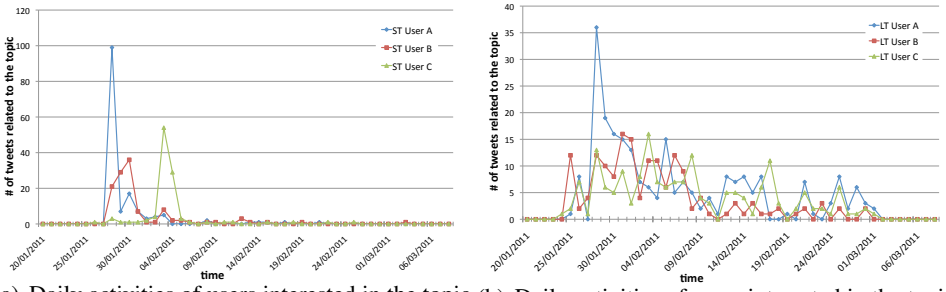


Figure 6.13: User adoption: number of new users per day who become interested in the Egyptian revolution.



(a) Daily activities of users interested in the topic for a short time period (*short-term adopters*).

(b) Daily activities of users interested in the topic for a long time period (*long-term adopters*).

Figure 6.14: Daily activities of users who are interested in the Egyptian revolution.

the Egyptian regime shut down the Internet on January 26th, about 300 users became interested into the protests on the "Friday of Rage", January 28th, and another 150 users took for the first time part in the Twitter discussions on the following day.

Having seen *when* individual users become for the first time interested in a topic, we were also interested for *how long* those users were interested in the topic. Figure 6.14 shows the amount of Twitter messages that selected users were posting on different days. The users whose tweeting activities on the topic of the Egyptian revolution are displayed in Figure 6.14(a) can be characterized as *short-term adopters* as they published tweets about the event for less than one week. It is interesting to see that the amount of messages these users posted about the topic is fairly high. For example, *ST User A*, who adopted the topic two days after the beginning of the

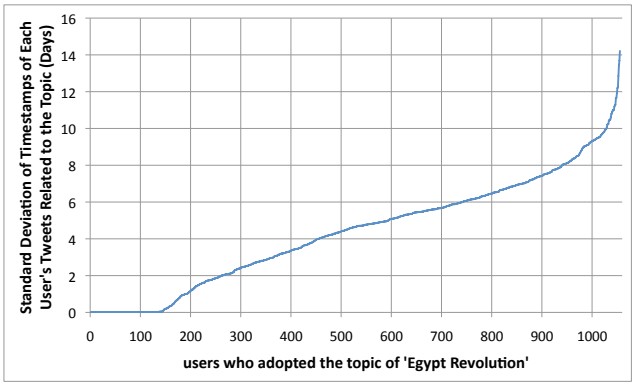


Figure 6.15: Standard Deviation of Timestamps of Related Tweets Posted by Each User

revolt, published almost 100 tweets about the revolution on a single day. Nevertheless, she quickly became disinterested. The interests of these three example users thus seem to change quickly. Hence, user modeling strategies that aim for capturing users' interests into topics have to adapt quickly as well.

Figure 6.14(b) displays the Twitter activities of three other users who were concerned with the Egyptian revolts for a long time period of more than one month and can therefore be considered as *long-term adopters*. All the three long-term adopters became interested into the topic at the very beginning of the revolt and can thus also be described as *early adopters*. In contrast, the short-term adopters characterized in Figure 6.14(a) are not among the first users who publish about the incidents in Egypt. In fact, for the Egyptian revolution it seems that there is a correlation between the time *when* a user adopts a topic and the duration during which the user is interested into the topic, i.e. early adopters overlap stronger with long-term adopters than with short-term adopters. Furthermore, the Twitter behavior of the short-term adopters regarding the Egyptian revolution is apparently more influenced by public trends than the behavior of the long-term adopters. For example, as depicted in Figure 6.14(a), *ST User A, B* and *C* show a peak after the riot on February 2nd that was entitled the *"Battle of the Camel"* and which was heavily discussed in social and mainstream news media. In contrast, the peaks of the long-term adopters, shown in Figure 6.14(b), happen much more frequently and also occur on days on which were not packed with epic events.

Figure 6.15 overviews the sample users who were interested in the Egyptian revolution with respect to the duration the different users expressed their interest into the topic on Twitter. In particular, it shows for each user the standard deviation of the timestamps of tweets that were related to the topic as similarly proposed by Huang et al. [93] who measure the temporal stability of hashtags. For Figure 6.15,

we apply standard deviation as follows.

$$\sigma(topic, user) = \sqrt{\frac{\sum_{k=1}^{N}(time(tweet_k) - \overline{time})^2}{N-1}} \tag{6.11}$$

Here, $time(tweet_k)$ is the timestamp of the k-th tweet published by the given *user* that refers to the given $topic$, $\overline{time}$ is the average timestamp of the user's tweets that relate to the topic and N is the overall number of tweets in which the user refers to the topic.

Figure 6.15 shows that for nearly 150 users the $\sigma(topic, user)$ is zero which means that those users just published one tweet that we could relate to the happenings in Egypt. Overall, for more than 75% of the users, the standard deviation of timestamps which specify when they published about the topic is less than one week. The fraction of long-term adopters for whom $\sigma(topic, user)$ is higher than ten days is with less than 2.5% rather low.

Findings

We have conducted a large scale analysis to study the evolution of user interests in trending topics. The findings of our analysis are summarized as follows.

1. Topics that are discussed on Twitter can be represented via the concepts that are referenced from the tweets that relate to the topic. Those concepts can be arbitrary entities such as persons, organizations or locations as well as cryptic hashtags like "#jan25". As different concepts may be of different relevance for a topic, it is desirable to weigh the concepts according to their importance for the topic.

2. Topics change over time: different concepts are of different importance for a given topic. For example, concepts such as *SMS* or *Vodafone* became important for the Egyptian revolution only for a short time when the government of Egypt shut down the Internet and took over the telecommunication network of Vodafone. Due to this event-like nature of a Twitter topic, it is helpful to compute the weight of a concept for a topic as a function of time.

3. The interests of individual users into a topic evolve differently over time in the context of the Egyptian revolution. Most users, who were interested in the topic, adopted the topic within a few days. Hence, the speed in which people adopt a topic on Twitter seems to be rather fast (cf. [106, 169]). However, the fraction of early adopters who become interested in an event on the day

the event happens is small. Moreover, the duration during which users are interested in an event-like topic differs clearly among the different users. In fact, we identified long-term adopters who are interested in a Twitter topic over a long period in time and short-term adopters who are concerned with a topic only for a short period in time and are rather driven by current trends.

6.4.2 Time-sensitive User Modeling for Personalized Recommendations

In Section 3.2.4, we introduced a time-sensitive weighting scheme which dampens the occurrence frequency according to the temporal distance between the topic occurrence time and the given timestamp (see Equation 3.6). Our hypothesis is that the time-sensitive strategy characterizes the actual demands and concerns of a user better than the non-time-sensitive baseline strategy. To investigate the above hypothesis, we deploy the user modeling strategies in a personalized recommender system. The recommender provides Web site recommendations to a user based on her user profile. We thus apply the Twitter-based user modeling strategies to personalize the Social Web experience of the users and point them to Web sites which are according to their profiles of interest in their current temporal context. We then study the following research questions.

1. How do semantic enrichment and time-sensitive weighting functions of the user modeling framework influence the performance of the recommender system?

2. Are there any correlations between characteristic patterns in the generated Twitter profiles and the gained recommendation quality? For example, how does the recommendation quality differ between users who have a tendency to be short-term or long-term adopters on a given topic?

Evaluation Methodology

We examine the influence of user modeling strategies on the performance of a recommender system that we developed for providing personalized Web site recommendations to the user. In particular those fresh Web sites that are referenced in Twitter messages (cf. [44, 58]). Recommending Web sites, which are posted on Twitter, is a non-trivial task as URLs. Our main goal is to analyze and compare the applicability of the different user modeling strategies in the context of the recommender system. We particularly analyze how the time-sensitive user modeling strategy influences personalization and performs in comparison to non-time-sensitive

variants. Therefore we apply the same recommendation algorithm introduced in Section 6.2.2.

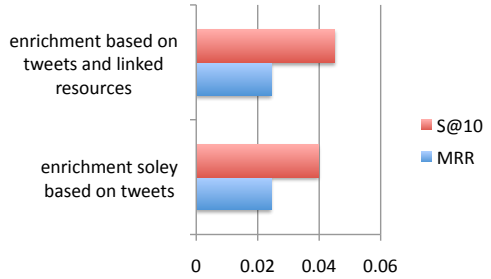
We compute personalized recommendations for each user of our sample on each day of our recommendation period which is given by the last ten days of January (Jan 20th - Jan 31st). Hence, our recommendation period overlaps with the beginning of the Egyptian revolution. However, the Web sites that are recommended to the users in this period may refer to any topic and are not necessarily related to the revolution in Egypt. The ground truth of URLs, which we consider as *relevant* for a specific user u on a particular day, is given by those Twitter messages which link to the corresponding Web site and which have been re-tweeted by u on that day. Following this evaluation strategy, we identified, on average, 24.5 relevant URLs for each of the 1619 sample users per day. The candidate set of URLs, which were published on a recommendation day, contained, on average, 24549 items.

Given the ground truth and candidate sets, we applied the different user modeling strategies together with the above algorithm and set of candidate items to compute fresh, personalized Web site recommendations for each user on each day. The user modeling strategies were only allowed to exploit tweets published before the start of the recommendation period. The quality of the recommendations was measured by means of $S@k$ (Success at rank k), which stands for the mean probability that a relevant item occurs within the top k of the ranking, and *MRR* (Mean Reciprocal Rank), which indicates at which rank the first item relevant to the user occurs on average. For Success@k, we will focus on S@10 as our recommendation system will list 10 Web site recommendations to a user. We tested statistical significance of our results with a two-tailed t-Test where the significance level was set to $\alpha = 0.01$ unless otherwise noted.

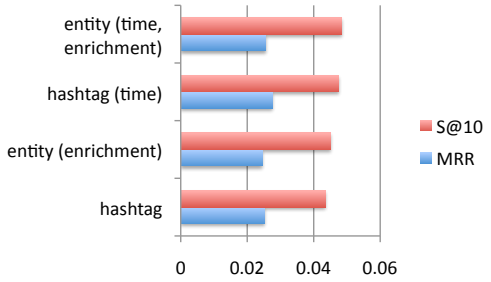
Results

Figure 6.16 summarizes the result of our recommendation experiment. In Figure 6.16(a), we first analyze the impact of the linkage enrichment provided by our user modeling framework. We observe that the recommendation quality is positively influenced by the enrichment component that follows the links in Twitter messages to also extract named entities from those Web pages. While the performance regarding MRR increases just slightly, S@10 improves by more than 15%. For the entity-based user modeling strategy, we thus apply the semantic enrichment method that exploits the links posted in Twitter messages also for the subsequent recommendation experiments.

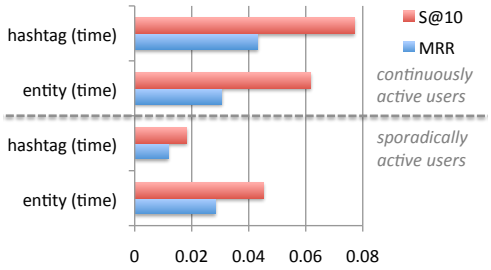
Figure 6.16(b) shows the performance of the entity-based and hashtag-based user modeling strategies and illustrates how the time-dependent weighting function



(a) Linkage Enrichment (entity-based profiles)



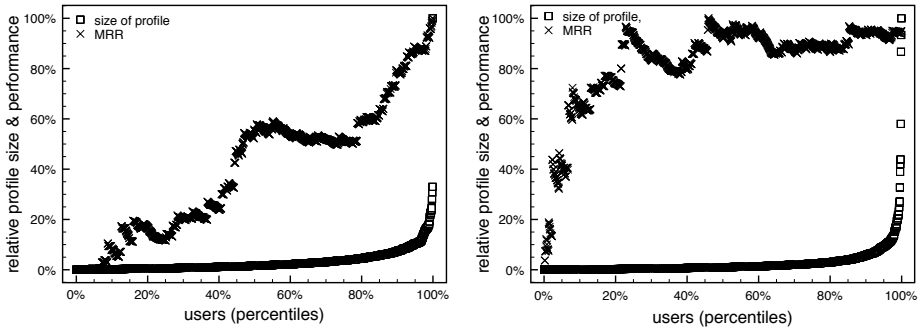
(b) Impact of Temporal Dynamics



(c) Different Types of Users

Figure 6.16: Comparison of user modeling strategies for supporting personalization.

(cf. Equation 3.6) influences the personalization quality. Regarding S@10, the entity-based user modeling strategy performs slightly better than the hashtag-based method (improvement: 5%). However, there is no significant difference in performance between entity-based and hashtag-based user modeling strategy. In contrast, the time-dependent weighting function increases the recommendation performance clearly. For the hashtag-based user modeling strategy, weighting the occurrence frequency according to the time for which a profile is demanded (*hashtag (time)*) improves the recommendation quality over the baseline strategy (*hashtag*) by 10.4% and 12% regarding S@10 and MRR respectively. We thus find first evidence for our hypothesis that the time-sensitive strategy characterizes the actual demands and



(a) Hashtag-based Profiles (b) Entity-based Profiles

Figure 6.17: Relation between size of profiles and quality of profiles for supporting personalization.

concerns of a user better than the non-time-sensitive baseline strategy.

Figure 6.16(c) illustrates the recommendation performance for different types of users: (i) people who are *continuously active* during our recommendation period and re-tweet at least one Web site on each day of the ten days (i.e. for each day there exist at least one relevant item to be recommended) and (ii) people who are *sporadically active* (on less than five days). As depicted in Figure 6.16(c), the recommendation performance is better for active users than for the sporadically active users. It is interesting to see that the hashtag-based version performs best for the continuously active users and rather fails for the sporadically active users for which the entity-based user modeling strategy performs best. Hence, it seems that for recommending Web sites on the Social Web, the interests of active users can be represented best via hashtags while the interests of sporadically active users are best modeled via the entity-based strategy.

Figure 6.17 further relates the recommendation quality with the size of entity-based and hashtag-based profiles. The size of a user's profile is measured by the number of distinct concepts that appear in a profile and is given relatively to the size of the biggest profile. The performance is measured via the mean reciprocal rank (MRR) and is also specified in a relative manner. Moreover, the MRR curves show the average performance for the corresponding x% of the users. For example, for those 20% of the users whose hashtag-based profiles are smaller than the profiles of the other 80% of the users, the recommendation quality is less than 20%. Figure 6.17(a) can thus be interpreted as follows: the bigger the hashtag-based profiles the better the recommendation.

For entity-based profiles, we observe different behavior, as depicted in Figure 6.17(b). The quality of recommendations computed based on entity-based pro-

files does not depend that strongly on the size of the profiles. In fact, it remains fairly stable for varying profile sizes.

Findings

We showed how the Twitter-based user modeling strategies can be applied in a recommender system to personalize the users' Social Web experience. The research questions raised at the beginning of this section can be answered as follows.

1. When determining the importance of concepts in a user profile, it is beneficial to weigh the concepts with respect to the point in time for which the profile is demanded. Those concepts which a user has been concerned with recently should be weighted higher than concepts which have not been referenced by the user for a long time. Moreover, we observed that entity-based user modeling performs best when extracting entities from both the Twitter messages and the Web resources, which are referenced from the corresponding Twitter message.

2. We also discovered remarkable correlations between the characteristics of the different types of user profiles and the resulting recommendation quality. When modeling users based on hashtags, the personalization performance correlates with the size of the hashtag-based profile: the bigger the profile, the better the performance. In contrast, personalization enabled via entity-based user modeling is highly independent from the size of a profile. Furthermore, we observed that for sporadically active users, which tend to be short-term adopters, the entity-based user modeling strategies, provided by our framework, perform much better than the hashtag-based strategies.

6.4.3 Synopsis

In this section, we analyzed the characteristics of topics discussed on Twitter and discovered that the representation of a topic changes over time: concepts related to a topic may gain or loose importance. For event-like topics, we identified different groups of users: *long-term adopters* join the discussion early and continuously contribute to the discussion while *short-term adopters* join the discussion later and participate just sporadically being influenced by public trends.

Based on this analysis, we introduced strategies that allow for incorporating those temporal characteristics into user profiles as well. We defined time-sensitive user modeling strategies (hashtag-based and entity-based) and evaluated these strate-

gies in context of a recommender system that provides personalized Web site recommendations. Our results prove the benefits of user modeling strategies that capture the temporal dynamics of a user's Twitter activities and reveal that semantic enrichment is particularly important for users who sporadically participate in the discussions on Twitter.

6.5 Domain-specific User Modeling on Twitter for Personalized Recommendations

In Chapter 3, we demonstrated that how our microblogging-based user modeling framework generates domain-specific user profiles. Give a stream of messages, GeniUS, which is a software library implemented based on our user modeling framework, allows for generating topics and user profiles that summarize the stream according domain- and application-specific needs which can be specified by the requesting party. Therefore, our user modeling framework can be applied in various application settings. In this section, we investigate the quality of user profiles that are adapted to six different domains for supporting various recommendation tasks. We tested statistical significance of our results with a two-tailed t-Test where the significance level was set to $\alpha = 0.01$ unless otherwise noted.

6.5.1 Analysis of Domain-Specific User Profile Construction

To understand the characteristics of user profiles constructed for supporting various application settings, we conducted an analysis on a large Twitter dataset. In our analysis, we investigate the characteristics of (i) complete Twitter-based profiles and (ii) six domain-specific types of profiles that were filtered using the semantic filtering method of GeniUS.

Data Collection

For our analysis, we monitored 73 Twitter users of the Social Handle Archive[3] (SoHArc) over a period of more than six months (from January 1st 2011 to July 7th). SoHArc lists profiles of researchers who are active in computer science and e-learning research in particular. Therefore, we ensured that there were no spam users in our dataset. Using the Twitter Streaming API via the *Item Fetcher* of GeniUS, we collected all public Twitter messages that these users published during the observation period. Overall, we thereby obtained 40,822 tweets. The seven most

[3] http://soharc.upb.de/

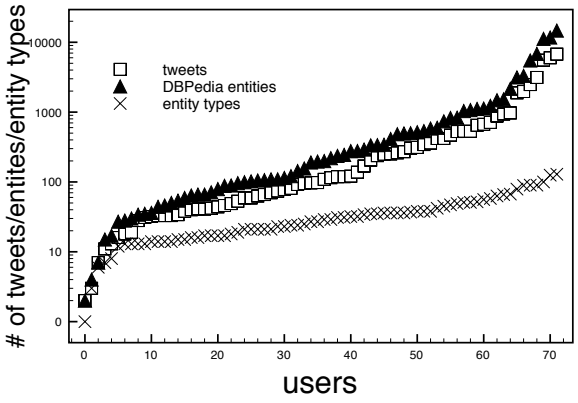
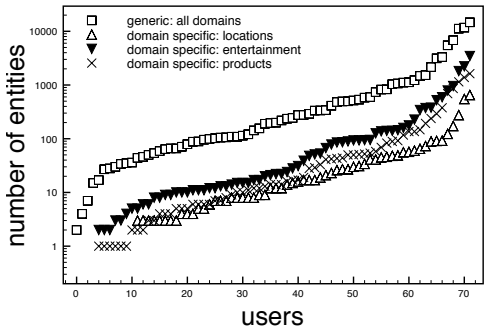


Figure 6.18: Number of tweets, DBpedia entities and entity types per user

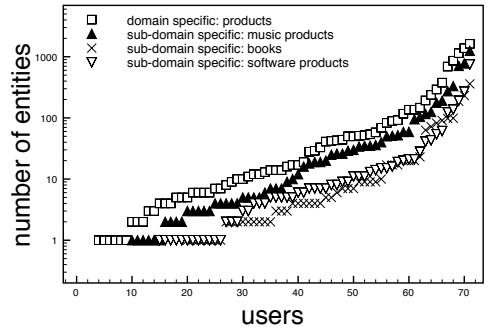
active users posted more than 1000 tweets while two users were almost inactive and published less than 10 Twitter messages (see Figure 6.18). We processed all Twitter messages with the semantic enrichment component of GeniUS which we configured so that DBpedia Spotlight was used as named entity recognition service as it allows for higher precision and recall than Alchemy or Zemanta (Mendes et al. report, for example, on precision of around 80% for disambiguating entities) [130]. Furthermore, we utilized the concept frequency as weighting scheme to compute weights for the entities of interest in the corresponding user profiles. Although all users are from the same computer science community, the topics about which they publish tweets show great variety as our analysis will reveal.

Results

In Figure 6.18, we plot the number of tweets and enriched DBpedia entities for each user. On average, each user published 567.0 Twitter messages and referred, according to the entity extraction module, to 1097.1 DBpedia entities. 59 of the users (82%) published more than 50 tweets during the observation period and also referred to more than 50 entities. And for each tweet, we extracted on average 1.9 entities. Each entity is identified by a unique URI. Therefore, we further retrieved – by resolving the URI – the types of the entities as specified in the corresponding DBpedia entry. The number of distinct types of entities to which a user refers to in her tweets is listed in Figure 6.18 as well. The average number of distinct types per user profile is 35.0 which indicates that there is a potential to generate different domain-specific profiles for a given user when categorizing entities of interest according to their types.



(a) Comparison of generic and domain specific user profile construction



(b) Comparison of domain and sub-domain specific user profile construction

Figure 6.19: Comparison of different strategies for user profile construction

To construct application domain based user profiles, we group the types of entities based on the DBpedia ontology [4] into several domains. In particular, we select three main domains for the analysis and further experiments: *location*, *entertainment* and *product*. Based on the hierarchies defined in the DBpedia ontology, we further derive three sub-domains from the product domain: *music products*, *books* and *software products*. In our evaluation, we classify items into these domains and test what kind of user profiles do best serve the domain-specific recommender system (recommender in the entertainment domain, book recommender etc.).

Using the semantic filter described in Section 3.3.2, we construct the (sub-) domain-specific user profiles. Figure 6.19 characterizes the corresponding profiles and shows the number of entities per user profile for the different types of profile construction strategies. In Figure 6.19(a), we compare the generic strategy, which utilizes all kinds of entities (no filtering), and the domain specific strategies which

[4] http://wiki.dbpedia.org/Ontology

filter the profiles so that they contain entities related to the location, entertainment and product domain respectively (semantic filtering). On average, there are 358.0 entities per user profile that belong to one of the three domains. Among them, 12.1% of the entities are categorized as locations (e.g. cities, countries and other places), 54.8% as being related to entertainment (e.g. sport, cultural events) and 33.1% as products (e.g. music albums, books, magazines, software). Some of these profiles are rather sparse. For example, for approximately 30 users, the location-based and product-related profiles contain less than 10 entities. The continuous difference between the size of the generic profiles and the domain-specific profiles indicates that all users reveal interests in different types of domains in their Twitter activities. Similarly, we observe in Figure 6.19(b) that also the domain-specific profiles related to products feature variety. When further filtering these profiles to obtain profiles that specify interests in music products, books and software products, one still obtains reasonably sized user profiles. Here, interests into music products can be inferred best. On average, 58% of the products mentioned by a user can be classified as music products (e.g. albums, songs) while 18.3% and 23.7% of the products are related to books (including newspapers and magazines) and software products respectively.

Our analysis thus shows that Twitter-based profiles reveal different types of interests of a user. Hence, there is potential to adapt Twitter-based profiles to different application domains. Figure 6.19 shows that we succeed in generating domain-specific profiles for the great majority of the users. The more specific the domain the smaller the profiles. Our hypothesis is that the stronger we adapt profiles to the given application domain – i.e. the more restrictive the filtering of the profiles – the better the performance of the corresponding application that consumes the profiles. In the subsequent section, we will investigate whether this hypothesis holds.

6.5.2 Evaluation of Domain-Specific User Profile Construction for Recommendation System

To test our hypothesis and to evaluate the quality of the Twitter-based user profiles that are created for different application domains, we apply the generated user profiles in different domain-specific recommendation systems and answer the following research questions:

1. Are the domain-specific user modeling strategies provided by GeniUS beneficial for supporting recommendation systems in different domains? For example, are the sparse – but more focused – user profiles more appropriate than the complete, unfiltered profiles?

Table 6.2: The average number of relevant items and candidate items for different recommendations.

application domain	***broad domains***			***product sub-domains***		
	entertainment	locations	products	music	books	software
average number of candidate items	1587.0	151.0	1207.0	756.0	237.0	254.0
average number of relevant items	70.0	13.4	49.6	40.9	16.4	20.3

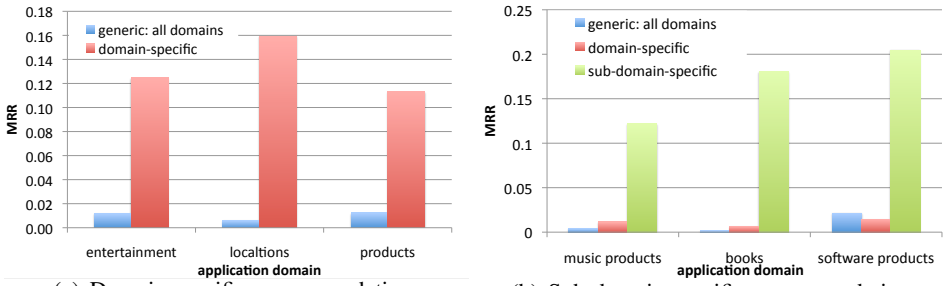
2. How does the performance vary between the different domains? For example, for what domains does the Twitter-based user modeling work best?

Experimental Setup

In our evaluation, we conduct tweet recommendation experiments to test the user profiles created for six different domains. The main goal of a domain-specific recommender is to recommend tweets to a user that are relevant to the given (sub-)domain and relevant to the user. We use the same content-based recommendation algorithm as introduced in Section 6.2 to recommends recommends items based on their cosine similarity with a given user, i.e. the more similar an item to a user the higher it will appear in the recommendation ranking.

For our experiments, we consider the last month of our observation period as the time frame for computing recommendations. For each of the six domains that we analyzed in Section 6.5.1, we deployed a recommendation system that used the algorithm described above as basis. Hence, the recommendation quality is solely influenced by the user modeling strategy for constructing user profiles. The ground truth of tweets which we consider as *relevant* to a specific user in a particular application domain is given by those messages that were actually posted by the user during the recommendation period and also contain at least one concept that belongs to the specified application domain. Hence, we remove all user information from the candidate tweets and try to assign the tweets to the right users by utilizing the user profiles that are constructed based on the tweets a user posted before the start of the recommendation period. The quality of recommendations is measured by means of *MRR* (Mean Reciprocal Rank) which indicates at what rank the first item relevant to the user occurs on average. We tested statistical significance of our results with a two-tailed t-Test where the significance level was set to $\alpha = 0.01$ unless otherwise noted.

The set of candidate items are those tweets that were published during the recommendation period and refer to at least one concept of the application domain of



(a) Domain specific recommendation (b) Sub-domain specific recommendation

Figure 6.20: Results of recommendation experiment

the recommendation system. The average number of relevant items per user and the number of candidate items in each application domain are listed in Table 6.2. For example, for the broader domains one can infer that the product domain is most challenging: the probability of randomly selecting a relevant item is 0.041 (= 49.6 / 1207.0) in contrast to 0.045 and 0.089 for the domains of entertainment and location respectively.

Results

The results of the recommendation experiments in the six different domains are summarized in Figure 6.20 and answer the research question raised at the beginning of this section. Figure 6.20(a) shows the quality of the recommendations in terms of MRR for the three broader domains: locations, entertainment and products. We compare the recommendation performances that were achieved based on the generic user modeling strategy, which does not make use of the semantic filtering functionality of GeniUS, and the domain-specific user modeling strategy, which filters out concepts from the profiles that are not related to the actual domain.

The domain-specific strategy consistently performs significantly better than the generic strategy ($p < 0.01$). While the domain-specific strategy achieves, on average, an MRR of 0.13 across the three different domains, the generic strategy performs poorly with 0.01. The domain-specific strategy produces with 0.16 regarding MRR the best results in context of the location-related recommendation system which is according to the proportion of relevant items per user the least challenging domain (see Table 6.2). In contrast, the generic strategy achieves only an MRR of 0.006. Within the context of product recommendations, the domain-specific profile construction method results in a ten times higher MRR and therefore outperforms the generic strategy clearly.

Similar results can be observed in the more narrow domains of music, book and software product recommendation systems. For the music and book domain, one can see that the recommendation quality increases the more specific the profiles are. In the music recommendation setting, the generic strategy fails with 0.004 regarding MRR, the product-specific profile improves the recommendation slightly (MRR: 0.01) and the profiles, which are specifically filtered for the music domain (*sub-domain-specific*), perform best and allow for an MRR of 0.12. The low performance of the generic user modeling strategy can be explained by noise that is introduced by entities that co-occur in tweets. For example, given that a user *u* tweets "Heading to Italy", GeniUS will infer that *u* has some interest into the concept *dbpedia:Italy*. The domain-specific strategy will filter out this interest when recommendations are computed in the music domain while the generic strategy will keep the information. Hence, given a candidate tweet such as "Justin Bieber concert in Italy #music", the domain-specific user modeling strategy will not cause this item to be recommended to *u* unless *u* showed interest into music of Justin Bieber in some of her other tweets that she published before the recommendation period. The generic user modeling strategy however will indicate to the recommender system that the given tweet might be of interest to the user because it mentions the concept *Italy* in which the user proved to be concerned with in the past. For the software domain this effect caused by the noise in the generic user profiles seems to be lower. However, again we observe that the sub-domain-specific user modeling strategy outperforms the other strategies clearly and allows for an MRR of 0.22.

Hence, regarding the research questions raised above, we can conclude that (1) domain-specific user modeling strategies provided by GeniUS allow for a tremendous improvement of the recommendation quality. Semantic filtering of the user profiles seems to remove noise and therefore allows us to adapt and optimize the user profile construction to the target domain. Moreover, we see that (2) the performance improvements are consistent throughout the different domains. The user modeling quality varies only slightly between the different domains. Furthermore, the performance does not seem to be influenced strongly by the size of the user profiles as a comparison of Figure 6.19 and Figure 6.20 reveals. With GeniUS, we thus succeed in generating domain-specific user profiles that allow for optimizing recommendation performance across different domains.

6.5.3 Synopsis

In this section, we demonstrated how our microblogging-based user modeling framework can be applied for supporting various application settings. Given (status) messages from microblogging services such as Twitter, GeniUS creates RDF-based representations that describe the semantic meaning of these messages. Based on the se-

mantically enriched user data, GeniUS provides different strategies for the creation of user interest profiles and provides means to semantically filter those profiles so that they adapt to a given application domain.

Our analysis based on Twitter status messages published by users during a period of several months showed that we succeed in generating profiles for different domains. To test the quality of the Twitter-based user profiles, we conducted recommendation experiments in six different domains and revealed that the domain-specific user modeling strategies, which filter user profiles and limit the concepts in the profiles to concepts related to the given application domain, allow clearly for the best performance.

6.6 Discussion

In Chapter 3, we presented the design space of our microblogging-based user modeling framework. By combining different design dimensions and design alternatives, our framework provides various user modeling strategies for constructing user profiles based on microblogging activities. In this chapter, we applied different user modeling strategies to construct microblogging-based user profiles that allow for supporting various personalization applications. We also explored how to integrate the popular trending topics into the user modeling process and investigated the temporal dynamics of trends and user profiles. Furthermore, we demonstrated how the user profile construction can be customized to serve a given application domain.

Given large Twitter datasets up to 10 million microposts, we analyzed the individual users' interests as well as public trends in the microblogging sphere to understand the characteristics of constructed trend and user profiles and explore the temporal dynamics of those profiles. Based on our analyses, we further evaluated the performance of user modeling strategies in context of various personalized recommender systems ranging from news recommendations to product recommendations in specific domains. With respect to the four experiments described at the beginning of this chapter, our main findings regarding the research questions are summarized in Table 6.3 and explained in detail as follows.

Analyzing user modeling for news recommendations Given a large Twitter dataset consisting of more than 2 million microblogging activities, we compared the quality of user profiles constructed with different topic modeling strategies in context of a personalized news recommender system. We investigated whether semantic enrichment with external Web resources and the consideration of temporal profile patterns can improve the recommendation quality. The findings can be summarized as follows.

Table 6.3: Overview on research questions investigated in this chapter.

Research question	Summary of findings
How do user modeling strategies impact the recommendation quality? Can the consideration of temporal patterns improve the accuracy of recommendations?	▶ The entity-based strategy enhances the variety of profiles and improves the recommendation quality. ▶ The semantic enrichment improves the performance of news recommendations ▶ The temporal patterns are beneficial for supporting personalized news recommendations
What is the impact of combining trend and user profiles on personalization?	▶ The time-sensitive weighting schemes succeed in modeling public trends. ▶ The combination of trend and user profiles improves the performance of news recommendations.
How are personal interests influenced by trends? What is the impact of time-sensitive weighting schemes on personalization?	▶ The individual users' interests into public trends evolve differently over time. ▶ The time-sensitive weighting schemes improve the accuracy of URL recommendations.
How do domain-specific user modeling strategies impact personalization?	▶ The adaptation of user profiles to a given domain improves the recommendation quality. ▶ The performance of recommender system varies between different application domains.

- We observed that the variety of entity-based profiles is much higher than the one of hashtag-based profiles (cf. *Hypothesis 1* in Table 3.6). For example, while the entity-based topic modeling strategies succeeds to construct for all users in our dataset, the hashtag-based strategy fails for 5.5% of users.
- The enrichment with semantics extracted from relevant external Web resources improved the accuracy of personalized news recommendations significantly (cf. *Hypothesis 2* in Table 3.6). For example, the exploitation both microposts and linked news articles for constructing user profiles improves MRR significantly.
- The results of our analysis revealed user profiles created based on microposts published on the weekends significantly differ from profiles created during the week. We further discovered that the consideration temporal profile patterns is beneficial to recommending news articles (cf. *Hypothesis 4* in Table 3.6).

Interweaving trend and user modeling for news recommendations We explored how our user modeling framework can be applied to model public trends based on microblogging data. We further analyzed and evaluated the quality of trend and user profiles in context of a trend-aware news recommender system. The main findings of this experiment can be summarized as follows.

- We analyzed trend and user profiles and discovered that certain parts of a user profiles change stronger over time than others. In particular, we observed that entities related to country or technology occur more constantly within profiles than entities related to movie or person.
- We also observed that some public trends are overloaded by entities that are constantly popular. The time-sensitive weighting function allows for filtering out those constantly popular entities and identifying trending entities for a specific period of time (cf. *Hypothesis 5* in Table 3.6).
- The results of our recommendation experiments showed the combination of trend and user profiles improves the the recommendation quality.

Analyzing temporal dynamics for URL recommendations By analyzing a concrete example trending topic, we illustrated how the individual users' interests into a trending topic change over time. Furthermore, we proposed time-sensitive weighting schemes to capture the temporal characteristics of user profiles and evaluated the quality of time-sensitive user modeling strategies in context of a URL recommender system. The main findings can be summarized as follows.

- The interests of individual users into trending topics evolve differently over time. We identified *long-term adopters* who are interested in topics over a long period of time and *short-term-adopter* who are interested in a topic only for a short period of time (cf. *Hypothesis 3* in Table 3.6).
- The results of URL recommendations illustrated that a time-sensitive weighting scheme, which weights the topics with respect to the point in time for which the profile is demanded, improves the accuracy of recommendations (cf. *Hypothesis 5* in Table 3.6).

Domain-specific user modeling for product recommendations We demonstrated how our user modeling framework allows for constructing user profiles that can be adapted to a given application domain. We conducted recommendation experiments in six different domains to evaluate the quality of the domain-specific user profiles. We summarize the main findings as follows.

- The results of our analysis revealed that we succeed in generating domain specific user profiles for the great majority of the users in our dataset.
- We also discovered that the adaptation of user profiles to a given application domain clearly improves the recommendation quality. The domain-specific profile construction strategy achieves ten times higher MRR than the generic user profile construction strategy in context of product recommender systems.

Chapter 7

Conclusion

With the advent of microblogging that becomes tangible in Social Web systems like Twitter, a new culture of participation penetrates the Web. The continuously growing amount of accessible microblogging data poses new challenges for user modeling and personalization. In this thesis, we tackled the challenges of inferring users' interests from microblogging activities and constructing semantically meaningful user profiles to support personalization in different contexts.

7.1 Summary of Contributions

Regarding the research questions that we identified in Section 2.3, the main findings and contributions of this thesis are summarized as follows.

Microblogging-based User Modeling Framework. As people discuss various topics on microblogging platforms, making sense of individual microblogging activities for user modeling is a non-trivial task. The following research questions were thus investigated.

- *How can users' personal interests be inferred from microblogging activities?*
- *How can we generate semantically meaningful user interest profiles that can be applied in different application domains?*

To answer the first question, we introduced a generic framework that allows for inferring individual users' interests from microblogging data and constructing semantically meaningful user profiles. We presented TweetUM, a framework for user modeling based on microblogging activities. It features

a variety of user modeling strategies for deducing users' personal interests from microblogging streams. We investigated different approaches for modeling topics of interests on microblogging platforms ranging from a hashtag-based strategy to more advanced strategies based on semantically meaningful concepts, for example, entities and categories that are extracted from microposts, and latent topics that are inferred via Latent Dirchlet Allocation. To overcome the shortness of micropost contents, our user modeling framework allows for enriching the semantics of microblogging activities by exploiting external resources such as external Web resources. The detailed description and evaluation of the semantic enrichment techniques was later discussed in Chapter 4. Furthermore, we introduced design alternatives for incorporating temporal constraints into the user modeling process. We demonstrated that the consideration of temporal constraints enabled to perform time period based filtering of microposts for inferring users' interests, and capture temporal patterns (e.g., weekday vs. weekend) which occur in user profiles. We also presented various weighting schemes for measuring the importance of topics of interests ranging from methods based on the occurrence of topics in microposts to time-sensitive weighting schemes that take into account temporal information when computing the weights of topics. We observed that the time-sensitive weighting schemes can better characterize the recent interests of a user than the non-time-sensitive weighting schemes

Regarding the second question, we explored strategies for modeling the semantics of microblogging activities and developed GeniUS, which is a software library implemented based on our user modeling framework. GeniUS consists of modules for collecting data from microblogging services, constructing semantically meaningful user profiles, and storing user profiles in RDF repositories for further processing. Following the design principles of Linked Data, user profiles are constructed using semantic concepts, which are identified via URIs, and therefore better support interoperability between different applications. Moreover, we demonstrated that GeniUS allows the third parties to customize the user modeling process for a given application domain. In summary, we presented in Chapter 3 a generic approach for generating user interest profiles based on microblogging data. We discussed flexible design choices for user modeling in the microblogging sphere and explored techniques that facilitate the process of user modeling, while further evaluation of our framework was done in Chapter 4-6.

Semantic Enrichment for Microblogging-based User Modeling. In order to construct valuable and meaningful user profiles that allow for better supporting external applications such as personalized recommender systems, in this thesis we presented approaches for constructing user profiles with rich semantics

and answered the following research questions.

- *How can we enrich the semantics of individual microblogging activities?*
- *How does the semantic enrichment impact the characteristics and quality of microblogging-based user profiles?*

We gave answers to the first question by introducing techniques that exploit two types of resources for the semantic enrichment of microblogging activities. First, we presented strategies for correlating microposts with external Web resources to enrich the semantics of microposts. Based on a large Twitter dataset, we evaluated different strategies for connecting microposts to related news articles on the Web. Our evaluation showed that by utilizing the entities extracted from both Twitter messages and news articles as well as the temporal information that indicates when Twitter messages and news articles were published, we succeeded in relating microposts and news articles with high precision and high coverage. Secondly, we introduced approaches for exploiting emotions expressed in microposts to further enrich the semantics of microblogging activities. We investigate how to construct emotion-based user profiles, which incorporate rich emotional facets into the user modeling process. We conducted classification experiments on a Twitter dataset to evaluate strategies for identifying different types of emotions from microposts and discovered that strategies based on semantically meaningful concepts (e.g., entities and WordNet concepts) achieved better performance than hashtag-based strategy in emotion classification experiments.

We answered the second question by conducting large-scale analyses of user profiles constructed with rich semantics. We saw that the exploitation of external Web resources (e.g., news articles that are related to microposts) allows for constructing user profiles with more detailed knowledge regarding user interests and therefore enhances the variety of user profiles. Furthermore, our analysis of emotion-based user profiles based on a large Twitter dataset revealed that the exploitation of emotions expressed in the microposts allows for capturing individual users' opinions about various topics and therefore delivers more insights into users' interests/concerns. For example, we observed that the variety of user profiles constructed for positive emotions is higher than the ones constructed for negative emotions.

Microblogging-based User Modeling for Culture-aware Analytics. To provide personalized services to individual users on microblogging platforms, it is important to understand user's preference and behavior patterns. Thus, we exploited various user characteristics to analyze microblogging behavior across

cultural groups and answered the research questions as follows.

- *How does the microblogging behavior vary between different cultural groups?*
- *Do differences in users microblogging behavior correlate with cultural theories in social sciences?*

Given the user modeling framework and the semantic enrichment of microposts, in Chapter 5 we answered the first question by analyzing user behavior across different microblogging platforms (Sina Weibo and Twitter) and cultural groups (Chinese and American users). We implemented functionality that allows for processing both Chinese and English microposts and presented the first large-scale comparative study of individual users' microblogging behavior. Our analyses revealed the key differences in microblogging behavior between Chinese and American users. We observed that American users are more triggered by hashtags and URLs when sharing and propagating information than Chinese users. The topics that Chinese users discussed on Sina Weibo are to a large extent related to locations and persons. In contrast, American users on Twitter discuss more about organizations in their posts and also utilize microblogging more often as part of their business compared to Chinese microbloggers who rather make use of microblogging services such Sina Weibo during their leisure time. Furthermore, Chinese users have a stronger tendency to publish and propagate messages that express positive sentiments than American users. Our results also revealed the differences in the temporal patterns of users' microblogging behavior between Chinese and American users.

Further, we answered the second research question by interpreting the findings of our comparative study with theories developed in social science research. We found clear correlation between our findings and the hypotheses about cultural commonalities and differences. For example, American users seem to be more eager to let their posts appear in the public discussion by more frequently including hashtags in their messages. This finding is possibly related to a higher demand of American users to profile themselves than Chinese users, which reflect the higher *individualism* in the Western cultures than in the Chinese culture. The positive nature of the information that people share on Sina Weibo again support the higher collectivism that is attributed to the Chinese culture. In summary, our finding and the interpretation of those findings deliver valuable insights for culture-aware user modeling.

Microblogging-based User Modeling for Personalized Recommendations. The large-scale analysis of users' microblogging behavior gave a good under-

standing of users' information needs and concerns. In this thesis, we further applied our user modeling framework in context of different personalized recommender systems and answered the research questions as follows.

- *How do the different user modeling strategies influence the quality of user profiles and the performance of personalized recommender system?*
- *What is the impact of incorporating trends and domain-specific knowledge into the user modeling process on the quality of personalized recommender systems?*

In Chapter 6, we analyzed and evaluated different user modeling strategies in context of various personalized recommender systems. The results of our analyses and recommendation experiments showed that the characteristics of user profiles as well as the quality of personalized recommendations are significantly influenced by different design dimensions and design alternatives. Among the proposed approaches for modeling topics of interests, the entity-based approach allows for constructing user profiles which have the highest variety and for achieving the best performance in news recommendation experiments. We observed that the semantic enrichment of microposts by exploiting external Web resources improves the quality of personalized news recommendations significantly. Our results also revealed the temporal patterns in user profiles. For example, the weekday user profiles significantly differ from the weekend user profiles. Furthermore, we showed that the consideration of temporal constraints for the construction of user profiles improves the performance of personalized recommender systems.

We presented approaches for modeling public trends on microblogging platform. By conducting large-scale analyses and examining some example trending topics, we observed that individual users' interests in public trends evolve over time. Further, we developed method to integrate trending topics into user profiles and showed that the integration improves the accuracy of personalized recommendations and therefore allows for better supporting trend-aware personalization.

Further, we showed that our user modeling framework allows for the construction of user profiles for a given application domain by utilizing external knowledge sources such as DBpedia. The results of recommendation experiments in six different domains revealed that the incorporation of domain-specific knowledge into the user modeling process improves the quality of personalized recommendations.

In summary, this thesis contributes to research on user modeling in the microblogging sphere as well as applications that exploit microblogging activities for

personalized recommendations. We developed a first generic framework for inferring individual users' interests from microblogging data. We introduced techniques for enriching the semantics of microblogging activities, which allows for constructing user profiles with semantically rich concepts. Given the user modeling framework, we analyzed microblogging behavior to understand individual users' information needs and discovered the key differences in users' microblogging behavior between different cultural groups. Moreover, we evaluated our microblogging-based user modeling framework in context of different recommender systems and showed the selection of design dimensions and design alternatives has significant impact on the characteristics of constructed user profiles as well as the quality of personalized recommendations.

7.2 Future Work

Based on the methods and findings presented in this thesis, we suggest the following recommendations for future work.

First, while the user modeling framework introduced in Chapter 3 focuses on mining users' interests based on the microposts published by those users, the *social relations* of users on microblogging platforms can be further investigated in order to obtain more relevant information for constructing user profiles. In particular, the exploitation of 'following' relationships could be beneficial to user modeling in the microblogging sphere. For example, Chen et al. exploited the Twitter messages published by a user as well as by other users (followees) that user follows to capture the user's interests [44]. However, Chen et al. constructed user profiles using a bag-of-words method, which only extract words from microposts. A possible direction for the future work is to apply the topic modeling strategies presented in this thesis to extract semantically meaningful topics from the microposts posted by a user's followees for modeling the interests of that user.

Second, as users leave a plethora of digital traces in various Social Web systems, *cross-system user modeling* can be a promising direction for further research. Abel et al. presented strategies that allow for linking and aggregating user profiles available in various Social Web systems (e.g., Facebook, Flickr, and Twitter) [8]. They discovered that the aggregation of user data distributed on the Social Web enhances the variety of user profiles and improves the quality of personalized recommendations significantly. The user modeling framework and semantic enrichment techniques presented in this thesis are mainly based on data collected from microblogging systems such as Twitter. User or usage data from other Social Web systems can be further exploited to cover different topics that a user discusses in

the specific systems. Moreover, the aggregation of user profiles might be beneficial for alleviating the sparsity problem in user modeling and personalized recommendations.

Third, given the understanding of users' preferences and behavior patterns across different cultural groups, the *culture-aware personalization* becomes a feasible research topic. In this thesis, we have conducted large-scale analyses based on microblogging data and delivered insights into the differences in users' microblogging behavior between Chinese and American users. Recently, researchers studied the adaptation of user interfaces to people who have different cultural backgrounds [155]. However, little research has been done on developing systems that take users' cultural backgrounds into account when generating personalized recommendations based on microblogging data. This is a topic that is well worthy of further investigation.

Finally, based on the domain-specific knowledge that is inferred from microblogging data, the *cross-domain user modeling in the context of collaborative filtering* can be further explored. Berkovsky et al. proposed an approach for importing and aggregating user information from other domains to construct user profiles and recommend items in a target domain using collaborative filtering techniques [18]. They showed that the aggregation of information across domains can make the recommender systems against sparsity problem and improve the accuracy of recommendations. While we have shown in this thesis that the incorporation of domain-specific knowledge allows for customizing the user profile construction for a given application domain and improves the quality of content-based recommendations in that domain, the impact of cross-domain user modeling strategies on the performance collaborative filtering recommender systems can be studied in the future.

Bibliography

[1] Fabian Abel. The benefit of additional semantics in folksonomy systems. In *Proceedings of the 2nd PhD workshop on Information and knowledge management*, PIKM '08, pages 49–56. ACM, 2008.

[2] Fabian Abel, Qi Gao, Geert-Jan Houben, and Ke Tao. Semantic enrichment of twitter posts for user profile construction on the social web. In *Proceedings of 8th Extended Semantic Web Conference*, ESWC '11, pages 375–389. Springer-Verlag, 2011.

[3] Fabian Abel, Qi Gao, Geert-Jan Houben, and Ke Tao. Supporting website: code, datasets and additional findings., 2011. http://wis.ewi.tudelft.nl/umap2011/.

[4] Fabian Abel, Qi Gao, Geert-Jan Houben, and Ke Tao. Analyzing temporal dynamics in twitter profiles for personalized recommendations in the social web. In *Proceedings of ACM 3rd International Conference on Web Science, Koblenz, Germany*, WebSci '11. ACM, 2011.

[5] Fabian Abel, Qi Gao, Geert-Jan Houben, and Ke Tao. Analyzing user modeling on twitter for personalized news recommendations. In *Proceedings of the 19th International Conference on User Modeling, Adaption and Personalization*, UMAP '11, pages 1–12. Springer-Verlag, 2011.

[6] Fabian Abel, Claudia Hauff, Geert-Jan Houben, Richard Stronkman, and Ke Tao. Semantics + filtering + search = twitcident. exploring information in social web streams. In *Proceedings of the 23rd ACM Conference on Hypertext and Social Media*, HT '12, pages 285–294. ACM 2012.

[7] Fabian Abel, Claudia Hauff, Geert-Jan Houben, and Ke Tao. Leveraging user modeling on the social web with linked data. In *Proceedings of the 12th International Conference on Web Engineering*, ICWE '12, pages 378–385. Springer-Verlag, 2012.

[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. Springer, 2013.

[9] Satyen Abrol and Latifur Khan. Twinner: understanding news queries with geo-content using twitter. In *Proceedings of the 6th Workshop on Geographic Information Retrieval*, GIR '10, pages 10:1–10:8. ACM, 2010.

[10] Gediminas Adomavicius and Alexander Tuzhilin. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *Knowledge and Data Engineering, IEEE Transactions on*, 17(6): 734–749. IEEE, 2005.

[11] Alexandre Ardichvili, Martin Maurer, Wei Li, Tim Wentling, and Reed Stuedemann. Cultural influences on knowledge sharing through online communities of practice. *Journal of Knowledge Management*, 10(1):94–107, 2006.

[12] Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. DBpedia: A nucleus for a web of open data. In *The semantic web*, pages 722–735. Springer, 2007.

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

[14] Marko Balabanović and Yoav Shoham. Fab: content-based, collaborative recommendation. *Communications of ACM*, 40(3):66–72, 1997.

[15] Dave Beckett. RDF/XML syntax specification (Revised). Technical report, World Wide Web Consortium (W3C), February 2004. URL http://www.w3.org/TR/REC-rdf-syntax/.

[16] David Beckett. N-triples: A line-based syntac for an rdf graph. Technical report, World Wide Web Consortium (W3C), April 2013. URL http://www.w3.org/TR/2013/NOTE-n-triples-20130409/.

[17] Robert Bell, Yehuda Koren, and Chris Volinsky. Modeling relationships at multiple scales to improve accuracy of large recommender systems. In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge discovery and data mining*, KDD '07, pages 95–104. ACM, 2007.

[18] Shlomo Berkovsky, Tsvi Kuflik, and Francesco Ricci. Cross-domain mediation in collaborative filtering. In *Proceedings of the 11th International*

Conference on User Modeling, UM '07, pages 355–359. Springer-Verlag, 2007.

[19] Tim Berners-Lee. WWW at 15 years: looking forward. In Allan Ellis and Tatsuya Hagino, editors, *Proceedings of the 14th International Conference on World Wide Web*, WWW '05, page 1. ACM, 2005.

[20] Tim Berners-Lee and Dan Connolly. Notation3 (N3): A readable RDF syntax. Technical report, World Wide Web Consortium (W3C), January 2008. URL http://www.w3.org/TeamSubmission/n3/.

[21] Tim Berners-Lee, James Hendler, and Ora Lassila. The Semantic Web. *Scientific American*, 284(5):34–43, 2001.

[22] Daniel Billsus and Michael J. Pazzani. User modeling for adaptive news access. *User Modeling and User-Adapted Interaction*, 10(2-3):147–180. Springer, 2000.

[23] Christian Bizer, Tom Heath, and Tim Berners-Lee. Linked data - the story so far. *International Journal on Semantic Web and Information Systems (IJSWIS)*, 5(3):1–22, 2009.

[24] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, March 2003. ISSN 1532-4435.

[25] U Bojārs, John G Breslin, Aidan Finn, and Stefan Decker. Using the semantic web for linking and reusing data across web 2.0 communities. *Web Semantics: Science, Services and Agents on the World Wide Web*, 6(1):page 21–28, 2008.

[26] Uldis Bojars and John G. Breslin. SIOC Core Ontology Specification. Namespace document, DERI, NUI Galway, http://rdfs.org/sioc/spec/, January 2009. URL http://rdfs.org/sioc/spec/.

[27] Uldis Bojars and John G. Breslin. SIOC Core Ontology Specification. Namespace document, DERI, NUI Galway, http://rdfs.org/sioc/spec/, January 2009. URL http://rdfs.org/sioc/spec/.

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

[29] 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*, UAI '98, pages 43–52. Morgan Kaufmann Publishers, 1998.

[30] John G. Breslin, Alexandre Passant, and Stefan Decker. *The social semantic web*. Springer, 1st edition, 2009.

[31] Dan Brickley and R.V. Guha. RDF Vocabulary Description Language 1.0: RDF Schema. W3C Recommendation, World Wide Web Consortium (W3C), February 2004. URL http://www.w3.org/TR/rdf-schema/.

[32] Dan Brickley and Libby Miller. FOAF Vocabulary Specification 0.91. Namespace document, FOAF Project, http://xmlns.com/foaf/0.1/, November 2007. URL http://xmlns.com/foaf/0.1/.

[33] Dan Brickley, Libby Miller, Toby Inkster, Yi Zeng, Yan Wang, Danica Damljanovic, Zhisheng Huang, Sheila Kinsella, John Breslin, and Bob Ferris. The Weighted Interests Vocabulary 0.5. Namespace document, Sourceforge, September 2010. URL http://purl.org/ontology/wi/core.

[34] Jeen Broekstra, Arjohn Kampman, and Frank van Harmelen. Sesame: A Generic Architecture for Storing and Querying RDF and RDF Schema. In I. Horrocks and J. Hendler, editors, *Proceedings of the First Internation Semantic Web Conference*, ISWC '02, ages 54–68. Springer, 2002.

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

[36] Peter Brusilovsky and Eva Millán. User models for adaptive hypermedia and adaptive educational systems. In *The adaptive web*, pages 3–53. Springer-Verlag, 2007.

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

[38] Francesca Carmagnola, Federica Cena, Luca Console, Omar Cortassa, Cristina Gena, Anna Goy, Ilaria Torre, Andrea Toso, and Fabiana Vernero. Tag-based user modeling for social multi-device adaptive guides. *User Modeling and User-Adapted Interaction*, 18(5):497–538, 2008.

[39] Francesca Carmagnola, Federica Cena, and Cristina Gena. User model interoperability: a survey. *User Modeling and User-Adapted Interaction*, 21(3): 285–331, 2011.

[40] Malu Castellanos, Umeshwar Dayal, Meichun Hsu, Riddhiman Ghosh, Mohamed Dekhil, Yue Lu, Lei Zhang, and Mark Schreiman. LCI: a social channel analysis platform for live customer intelligence. In *Proceedings of the 2011 ACM SIGMOD International Conference on Management of Data*, SIGMOD '11, pages 1049–1058. ACM, 2011.

[41] Steve Cayzer. Semantic blogging and decentralized knowledge management. *Communications of ACM*, 47(12):47–52, 2004.

[42] Meeyoung Cha, Hamed Haddadi, Fabrcio Benevenuto, and P. Krishna Gummadi. Measuring user influence in twitter: The million follower fallacy. In *Proceedings of the Fourth International Conference on Weblogs and Social Media*, ICWSM '10. The AAAI Press, 2010.

[43] Milen Chechev and Petko Georgiev. A multi-view content-based user recommendation scheme for following users in twitter. In *Proceedings of the 4th International Conference on Social Informatics*, SocInfo'12, pages 434–447. Springer-Verlag, 2012.

[44] 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 28th International Conference on Human Factors in Computing Systems*, CHI '10, pages 1185–1194. ACM, 2010.

[45] Li Chen and Ho Keung Tsoi. Analysis of user tags in social music sites: Implications for cultural differences. In *Extended Abstracts of ACM Conference on Computer Supported Cooperative Work*, CSCW '11. ACM, 2011.

[46] Zhiyuan Cheng, James Caverlee, and Kyumin Lee. You are where you tweet: a content-based approach to geo-locating twitter users. In *Proceedings of the 19th ACM International Conference on Information and knowledge management*, CIKM '10, pages 759–768. ACM, 2010

[47] Marc Cheong and Vincent Lee. Integrating web-based intelligence retrieval and decision-making from the twitter trends knowledge base. In *Proceedings of the 2nd ACM workshop on Social web search and mining*, SWSM '09, pages 1–8. ACM, 2009.

[48] Henriette Cramer, Vanessa Evers, Satyan Ramlal, Maarten Someren, Lloyd Rutledge, Natalia Stash, Lora Aroyo, and Bob Wielinga. The effects of transparency on trust in and acceptance of a content-based art recommender. *User Modeling and User-Adapted Interaction*, 18(5):455–496, 2008.

[49] Abhinandan S. Das, Mayur Datar, Ashutosh Garg, and Shyam Rajaram. Google news personalization: scalable online collaborative filtering. In *Proceedings of the 16th International Conference on World Wide Web*, WWW '07, pages 271–280. ACM, 2007.

[50] Paul De Bra, Ad Aerts, Bart Berden, Barend de Lange, Brendan Rousseau, Tomi Santic, David Smits, and Natalia Stash. AHA! the adaptive hypermedia architecture. In *Proceedings of the 4th ACM Conference on Hypertext and Hypermedia*, HT '03, pages 81–84. ACM, 2003.

[51] Munmun De Choudhury, Scott Counts, and Michael Gamon. Not all moods are created equal! exploring human emotional states in social media. In *Proceedings of the 6th Internaltional AAAI Conference on Weblogs and Social Media*, ICWSM '12. The AAAI Press, 2012.

[52] Delicious. http://www.delicious.com.

[53] Mukund Deshpande and George Karypis. Item-based top-n recommendation algorithms. *ACM Trans. Inf. Syst.*, 22(1):143–177, 2004.

[54] Christian Desrosiers and George Karypis. A comprehensive survey of neighborhood-based recommendation methods. In *Recommender Systems Handbook*, pages 107–144. Springer, 2011.

[55] Nicholas A. Diakopoulos and David A. Shamma. Characterizing debate performance via aggregated twitter sentiment. In *Proceedings of the 28th International Conference on Human factors in computing systems*, CHI '10, pages 1195–1198. ACM, 2010.

[56] Vania Dimitriva, Royce Neagle, Sirisha Bajanki, Lydia Lau, and Roger Boyle. Awesome computing: using corpus data to tailor a community environment for dissertation writing. In *Proceedings of the 10th International Conference on Intelligent Tutoring Systems - Volume Part II*, ITS '10, pages 443–443. Springer-Verlag, 2010

[57] Yi Ding and Xue Li. Time weight collaborative filtering. In *Proceedings of the 14th ACM International Conference on Information and Knowledge Management*, CIKM '05, pages 485–492. ACM, 2005.

[58] 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. ACM, 2010.

[59] Miles Efron. Hashtag retrieval in a microblogging environment. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in information retrieval*, SIGIR '10, pages 787–788. ACM, 2010.

[60] Paul Ekman. Universals and cultural differences in facial expressions of emotion. In *Nebraska Symposium on Motivation*. University of Nebraska Press, 1972.

[61] Facebook. http://www.facebook.com.

[62] C. Fellbaum and I. NetLibrary. *WordNet: an electronic lexical database*. MIT Press USA, 1998.

[63] Claudiu S. Firan, Wolfgang Nejdl, and Raluca Paiu. The benefit of using tag-based profiles. In *Proceedings of the 2007 Latin American Web Conference*, LA-WEB '07, pages 32–41. IEEE, 2007.

[64] Devin Gaffney. #iranElection: quantifying online activism. In *Proceedings of International Conference on Web Science*, WebSci '10. 2010.

[65] Qi Gao, Junwei Yan, and Min Liu. A semantic approach to recommendation system based on user ontology and spreading activation model. In *Proceedings of 2008 International Conference on Network and Parallel Computing*, NPC '08, pages 488–492. IEEE, 2008.

[66] Qi Gao, Fabian Abel, and Geert-Jan Houben. GeniUS: Generic user modeling library for the social semantic web. In *Proceedings of Joint International Semantic Technology Conference*, JIST '11, page 160–175. Springer, 2011.

[67] Qi Gao, Fabian Abel, Geert-Jan Houben, and Ke Tao. Interweaving trend and user modeling for personalized news recommendations. In *Proceedings of the 2011 International Conference on Web Intelligence*, WI '11. IEEE, 2011.

[68] Qi Gao, Fabian Abel, Geert-Jan Houben, and Yong Yu. A comparative study of users' microblogging behavior on sina weibo and twitter. In *Proceedings of the 20th International Conference on User Modeling, Adaptation, and Personalization*, UMAP '12, pages 88–101. Spinger-Verlag, 2012.

[69] Qi Gao, Fabian Abel, Geert-Jan Houben, and Yong Yu. Information propagation cultures on sina weibo and twitter. In *Proceedings of the ACM Web Science Conference 2012*, WebSci'12. ACM, 2012.

[70] Ruth Garcia-Gavilanes, Daniele Quercia, and Alejandro Jaimes. Cultural dimensions in twitter: Time, individualism and power. In *Proceedings of the*

International Conference on Weblogs and Social Media, ICWSM '13. The AAAI Press, 2013.

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

[72] Anna Lisa Gentile, Vitaveska Lanfranchi, Suvodeep Mazumdar, and Fabio Ciravegna. Extracting semantic user networks from informal communication exchanges. In *Proceedings of the 10th International Conference on The Semantic Web - Volume Part I*, ISWC '11, pages 209–224. Springer-Verlag, 2011.

[73] Gianluigi Gentili, Alessandro Micarelli, and Filippo Sciarrone. Infoweb: An adaptive information filtering system for the cultural heritage domain. *Applied Artificial Intelligence*, 17(8-9):715–744, 2003.

[74] M. Rami Ghorab, Dong Zhou, Alexander OConnor, and Vincent Wade. Personalised information retrieval: survey and classification. *User Modeling and User-Adapted Interaction*, pages 1–63. Springer, 2012.

[75] Saptarshi Ghosh, Naveen Sharma, Fabricio Benevenuto, Niloy Ganguly, and Krishna Gummadi. Cognos: crowdsourcing search for topic experts in microblogs. In *Proceedings of the 35th international ACM Conference on Research and Development in Information Retrieval*, SIGIR '12, pages 575–590. ACM, 2012.

[76] Alec Go, Richa Bhayani, and Lei Huang. Twitter sentiment classification using distant supervision. Technical report, Stanford University, 2009.

[77] Namrata Godbole, Manjunath Srinivasaiah, and Steven Skiena. Large-scale sentiment analysis for news and blogs. In *Proceedings of the International Conference on Weblogs and Social Media*, ICWSM '07. The AAAI Press, 2007.

[78] Jen Golbeck and Derek L. Hansen. Computing Political Preference among Twitter Followers. In *Proceedings of the 29th International Conference on Human Factors in Computing Systems*, CHI '11. ACM, 2011.

[79] Jennifer Golbeck, Cristina Robles, Michon Edmondson, and Karen Turner. Predicting personality from twitter. In *Proceedings of the 2011 International Conference on Social Computing*, SocialCom '11, pages 149–156. IEEE, 2011.

[80] Google+. http://plus.google.com.

[81] Derek Greene, Gavin Sheridan, Barry Smyth, and Pádraig Cunningham. Aggregating content and network information to curate twitter user lists. In *Proceedings of the 4th ACM RecSys workshop on Recommender systems and the social web*, RSWeb '12, pages 29–36. ACM, 2012.

[82] Thomas Gruber. Toward principles for the design of ontologies used for knowledge sharing. *International Journal of Human Computer Studies*, 43 (5):907–928, 1995.

[83] Tom Gruber. Collective knowledge systems: Where the social web meets the semantic web. *Web Semantics: Science, Services and Agents on the World Wide Web*, 6(1):4–13, 2008.

[84] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H Witten. The weka data mining software: an update. *ACM SIGKDD Explorations Newsletter*, 11(1):10–18, 2009.

[85] John Hannon, Mike Bennett, and Barry Smyth. Recommending twitter users to follow using content and collaborative filtering approaches. In *Proceedings of the fourth ACM conference on Recommender systems*, RecSys '10, pages 199–206. ACM, 2010.

[86] Claudia Hauff, Marcel Berthold, Geert-Jan Houben, Christina M. Steiner, and Dietrich Albert. Tweets reveal more than you know: a learning style analysis on twitter. In *Proceedings of the 7th European Conference on Technology Enhanced Learning*, EC-TEL '12, pages 140–152. Springer-Verlag, 2012.

[87] Tom Heath and Enrico Motta. Revyu.com: A reviewing and rating site for the web of data. In *Proceedings of the 6th International Semantic Web and 2nd Asian Conference on Asian Semantic Web Conference*, ISWC'07/ASWC'07, pages 895–902. Springer-Verlag, 2007.

[88] Brent Hecht, Lichan Hong, Bongwon Suh, and Ed H. Chi. Tweets from justin bieber's heart: The dynamics of the "location" field in user profiles. In *Proceedings of the 29th International Conference on Human Factors in Computing Systems*, CHI' 11, page 237–246. ACM, 2011.

[89] Dominik Heckmann, Tim Schwartz, Boris Brandherm, Michael Schmitz, and Margeritta von Wilamowitz-Moellendorff. Gumo - the general user model ontology. In *Proceedings of the 10th International Conference on User Modeling*, UM '05, pages 428–432. Springer, 2005.

[90] Ivan Herman, Guy Melançon, and M. Scott Marshall. Graph visualization and navigation in information visualization: A survey. *IEEE Transaction on Visualization and Computer Graphics*, 6 (1):24–43, 2000.

[91] Geert Hofstede and Gert Jan Hofstede. *Cultures and Organizations: Software of the Mind*. McGraw-Hill, 2005.

[92] Liangjie Hong and Brian D. Davison. Empirical study of topic modeling in twitter. In *Proceedings of the First Workshop on Social Media Analytics*, SOMA '10, pages 80–88. ACM, 2010.

[93] Jeff Huang, Katherine M. Thornton, and Efthimis N. Efthimiadis. Conversational tagging in twitter. In *Proceedings of the 21st ACM conference on Hypertext and hypermedia*, HT '10, pages 173–178. ACM, 2010.

[94] Ashutosh Jadhav, Hemant Purohit, Pavan Kapanipathi, Pramod Ananthram, Ajith Ranabahu, Vinh Nguyen, Pablo N. Mendes, Alan Gary Smith, Michael Cooney, , and Amit Sheth. Twitris 2.0 : Semantically empowered system for understanding perceptions From social data. In *Semantic Web Challenge*, 2010.

[95] Anthony Jameson, Cécile Paris, and Carlo Tasso. Reader's guide. In *6th International Conference on User Modeling*, 1997.

[96] Akshay Java, Xiaodan Song, Tim Finin, and Belle Tseng. Why we twitter: understanding microblogging usage and communities. In *Proceedings of the 1st Workshop on Web Mining and Social Network Analysis*, WebKDD/SNA-KDD '07, pages 56–65. ACM, 2007.

[97] David R Karger and Dennis Quan. What would it mean to blog on the semantic web? In *The Semantic Web*, pages 214–228. Springer, 2004.

[98] Hak Lae Kim, Simon Scerri, John G Breslin, Stefan Decker, and Hong Gee Kim. The state of the art in tag ontologies: a semantic model for tagging and folksonomies. In *International Conference on Dublin Core and Metadata Applications*, pages 128–137, 2008.

[99] Alfred Kobsa. Generic user modeling systems. *User Modeling and User-Adapted Interaction*, 11(1-2):49–63, 2001.

[100] Alfred Kobsa. Privacy-enhanced personalization. *Communications of ACM*, 50(8):24–33, 2007.

[101] 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*, WSDM '10, pages 441–450. ACM, 2010.

[102] Jonathan Koren, Yi Zhang, and Xue Liu. Personalized interactive faceted search. In *Proceedings of the 17th International Conference on World Wide Web*, WWW '08, pages 477–486. ACM, 2008.

[103] Yehuda Koren, Robert Bell, and Chris Volinsky. Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37, 2009.

[104] Essam Kosba, Vania Dimitrova, and Roger Boyle. Adaptive feedback generation to support teachers in web-based distance education. *User Modeling and User-Adapted Interaction*, 17(4):379–413, 2007.

[105] Markus Krötzsch, Denny Vrandečić, and Max Völkel. Semantic mediawiki. In *The Semantic Web*, pages 935–942. Springer, 2006.

[106] 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*, WWW '10, pages 591–600. ACM, 2010.

[107] Su Mon Kywe, Tuan-Anh Hoang, Ee-Peng Lim, and Feida Zhu. On recommending hashtags in twitter networks. In *Proceedings of the 4th International Conference on Social Informatics*, SocInfo '12, pages 337–350. Springer-Verlag, 2012.

[108] Su Mon Kywe, Ee-Peng Lim, and Feida Zhu. A survey of recommender systems in twitter. In *Proceedings of the 4th International Conference on Social Informatics*, SocInfo '12, pages 420–433. Springer-Verlag, 2012.

[109] David Laniado and Peter Mika. Making sense of twitter. In *Proceedings of the 9th International Semantic Web Conference*, ISWC '10, pages 42–51. Springer, 2010.

[110] Changhyun Lee, Haewoon Kwak, Hosung Park, and Sue Moon. Finding influentials based on the temporal order of information adoption in twitter. In *Proceedings of the 19th International Conference on World wide web*, WWW '10, pages 1137–1138. ACM, 2010.

[111] Janette Lehmann, Bruno Gonçalves, José J. Ramasco, and Ciro Cattuto. Dynamical classes of collective attention in twitter. In *Proceedings of the 21st*

International Conference on World Wide Web, WWW '12, pages 251–260. ACM, 2012.

[112] Kristina Lerman and Rumi Ghosh. Information contagion: an empirical study of spread of news on digg and twitter social networks. In *Proceedings of 4th International Conference on Weblogs and Social Media*, ICWSM '10. The AAAI Press, 2010.

[113] Anton Leuski and James Allan. Interactive information retrieval using clustering and spatial proximity. *User Modeling and User-Adapted Interaction*, 14(2-3):259–288, 2004.

[114] Lei Li, Li Zheng, and Tao Li. Logo: a long-short user interest integration in personalized news recommendation. In *Proceedings of the 5th ACM conference on Recommender systems*, RecSys '11, pages 317–320. ACM, 2011.

[115] Huizhi Liang, Yue Xu, Dian Tjondronegoro, and Peter Christen. Time-aware topic recommendation based on micro-blogs. In *Proceedings of the 21st ACM International Conference on Information and knowledge management*, CIKM '12, pages 1657–1661. ACM, 2012.

[116] Greg Linden, Brent Smith, and Jeremy York. Amazon. com recommendations: Item-to-item collaborative filtering. *Internet Computing, IEEE*, 7(1): 76–80, 2003.

[117] Jiahui Liu, Peter Dolan, and Elin Rønby Pedersen. Personalized news recommendation based on click behavior. In *Proceedings of the 14th International Conference on Intelligent user interfaces*, IUI '10, pages 31–40. ACM, 2010.

[118] Rui Long, Haofen Wang, Yuqiang Chen, Ou Jin, and Yong Yu. Towards effective event detection, tracking and summarization on microblog data. In *Proceedings of the 12th International Conference on Web-age Information Management*, WAIM'11, pages 652–663. Springer-Verlag, 2011.

[119] Pasquale Lops, Marco de Gemmis, and Giovanni Semeraro. Content-based recommender systems: State of the art and trends. In *Recommender Systems Handbook*, pages 73–105. Springer, 2011.

[120] Sofus Macskassy and Matthew Michelson. Why do people retweet? anti-homophily wins the day! In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*, ICWSM '11. The AAAI Press, 2011.

[121] Thomas Mandl. Comparing chinese and german blogs. In *Proceedings of the 20th ACM Conference on Hypertext and Hypermedia*, HT '09, pages 299–308. ACM, 2009.

[122] Cameron Marlow, Mor Naaman, Danah Boyd, and Marc Davis. Ht06, tagging paper, taxonomy, flickr, academic article, to read. In *Proceedings of the seventeenth conference on Hypertext and hypermedia*, HT '06, pages 31–40. ACM, 2006.

[123] Nathalie Mathe and James R Chen. A user-centered approach to adaptive hypertext based on an information relevance model. In *Proceedings of 4th International Conference on User Modeling*, 1994.

[124] Adam Mathes. Folksonomies-cooperative classification and communication through shared metadata. *Computer Mediated Communication*, 47(10), 2004.

[125] Michael Mathioudakis and Nick Koudas. Twittermonitor: trend detection over the twitter stream. In *Proceedings of the 2010 International Conference on Management of Data*, SIGMOD '10, pages 1155–1158. ACM, 2010.

[126] Thomas Vander Wal. Folksonomy. *http://vanderwal.net/folksonomy.html*, July 2007.

[127] Deborah L. McGuinness and Frank van Harmelen. Owl web ontology language overview. Technical report, World Wide Web Consortium (W3C), February 2004. URL http://www.w3.org/TR/owl-features/.

[128] Bhaskar Mehta. *Cross system personalization: Enabling personalization across multiple systems*. PhD thesis, Universität Duisburg-Essen, Fakultät für Ingenieurwissenschaften» Ingenieurwissenschaften-Campus Duisburg» Abteilung Informatik und Angewandte Kognitionswissenschaft» Informationssysteme, 2008.

[129] Pablo N. Mendes, Alexandre Passant, and Pavan Kapanipathi. Twarql: tapping into the wisdom of the crowd. In *Proceedings of the 6th International Conference on Semantic Systems*, I-SEMANTICS '10, pages 45:1–45:3. ACM, 2010.

[130] Pablo N. Mendes, Max Jakob, Andrés García-Silva, and Christian Bizer. Dbpedia spotlight: Shedding light on the web of documents. In *Proceedings of the 7th International Conference on Semantic Systems*, I-SEMANTICS '11. ACM, 2011.

[131] Marcelo Mendoza, Barbara Poblete, and Carlos Castillo. Twitter under crisis: Can we trust what we rt? In *Proceedings of the 1st Workshop on Social Media Analytics*. SOMA '10. ACM, 2010.

[132] Matthew Michelson and Sofus A. Macskassy. Discovering users' topics of interest on twitter: a first look. In *Proceedings of the 4th Workshop on Analytics for Noisy Unstructured Text Data*, AND '10, pages 73–80. ACM, 2010.

[133] Stuart E Middleton, David C De Roure, and Nigel R Shadbolt. Capturing knowledge of user preferences: ontologies in recommender systems. In *Proceedings of the 1st International Conference on Knowledge Capture*, K-CAP '01, pages 100–107. ACM, 2001.

[134] Peter Mika. Ontologies are us: A unified model of social networks and semantics. *Web Semantic*, 5(1):5–15, 2007.

[135] Alexander Mikroyannidis. Toward a social semantic web. *Computer*, 40(11): 113–115, 2007.

[136] Vuk Milicic and Serbia Faviki. Case study: Semantic tags. *W3C Semantic Web Case Studies and Use Cases (December 2008) http://www. w3. org/2001/sw/sweo/public/Use-Cases/Faviki*, 2008.

[137] George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.

[138] Kanika Narang, Seema Nagar, Sameep Mehta, L. V. Subramaniam, and Kuntal Dey. Discovery and analysis of evolving topical social discussions on unstructured microblogs. In *Proceedings of the 35th European Conference on Advances in Information Retrieval*, ECIR'13, pages 545–556. Springer-Verlag, 2013.

[139] Richard Newman, Danny Ayers, and Seth Russell. Tag ontology, December 2005.

[140] Tim Oreilly. What is web 2.0: Design patterns and business models for the next generation of software. *Communications & strategies*, page 17, 2007.

[141] Ashley M. Oudenne, Youngmoo E. Kim, and Douglas S. Turnbull. Meerkat: exploring semantic music discovery using personalized radio. In *Proceedings of the International Conference on Multimedia information retrieval*, MIR '10, pages 429–432. ACM, 2010.

[142] Bo Pang and Lillian Lee. Opinion mining and sentiment analysis. *Found. Trends Inf. Retr.*, 2(1-2):1–135, 2008.

[143] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - Volume 10*, EMNLP '02, pages 79–86. ACL 2002.

[144] Alexandre Passant. Using ontologies to strengthen folksonomies and enrich information retrieval in weblogs. In *Proceedings of International Conference on Weblogs and Social Media*, ICWSM '07. The AAAI Press, 2007.

[145] Alexandre Passant and Philippe Laublet. Meaning of a tag: A collaborative approach to bridge the gap between tagging and linked data. In *Proceedings of the WWW 2008 Workshop Linked Data on the Web*, LDOW '08, 2008.

[146] Alexandre Passant, Tuukka Hastrup, Uldis Bojars, and John Breslin. Microblogging: A Semantic Web and Distributed Approach. In Christian Bizer, Sören Auer, Gunnar Aastrand Grimnes, and Tom Heath, editors, *Proceedings of the the 4th Workshop Scripting for the Semantic Web*, SFSW '08, volume 368. CEUR-WS.org, 2008.

[147] Michael Pazzani and Daniel Billsus. Learning and revising user profiles: The identification of interesting web sites. *Machine learning*, 27(3):313–331, 1997.

[148] 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*, RecSys '09, pages 385–388. ACM, 2009.

[149] Eric Prud'hommeaux and Andy Seaborne. SPARQL Query Language for RDF. W3c recommendation, W3C, January 2008. URL http://www.w3.org/TR/rdf-sparql-query/.

[150] Matthew Purver and Stuart Battersby. Experimenting with distant supervision for emotion classification. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '12, pages 482–491. ACL, 2012.

[151] Yan Qu, Chen Huang, Pengyi Zhang, and Jun Zhang. Microblogging after a major disaster in china: a case study of the 2010 yushu earthquake. In *Proceedings of the ACM 2011 Conference on Computer Supported Cooperative Work*, CSCW '11, pages 25–34. ACM, 2011.

[152] Daniel Ramage, Susan T. Dumais, and Daniel J. Liebling. Characterizing microblogs with topic models. In *Proceedings of the 4th International Conference on Weblogs and Social Media*, ICWSM '10. The AAAI Press, 2010.

[153] Tye Rattenbury, Nathaniel Good, and Mor Naaman. Towards automatic extraction of event and place semantics from Flickr tags. In *Proceedings of the 30th International ACM SIGIR Conference on Information Retrieval*, SIGIR '07, pages 103–110. ACM, 2007.

[154] Katharina Reinecke and Abraham Bernstein. Tell me where you've lived, and i'll tell you what you like: Adapting interfaces to cultural preferences. In *Proceedings of the 17th International Conference on User Modeling, Adaptation, and Personalization*, UMAP '09, pages 185–196. Springer-Verlag, 2009.

[155] Katharina Reinecke and Abraham Bernstein. Improving performance, perceived usability, and aesthetics with culturally adaptive user interfaces. *ACM Transaction on Computer Human Interaction*, 18(2):8:1–8:29, 2011.

[156] Katharina Reinecke, Minh Khoa Nguyen, Abraham Bernstein, Michael Näf, and Krzysztof Z. Gajos. Doodle around the world: online scheduling behavior reflects cultural differences in time perception and group decision-making. In *Proceedings of the 2013 conference on Computer supported cooperative work*, CSCW '13, pages 45–54. ACM, 2013.

[157] Paul Resnick and Hal R. Varian. Recommender systems. *Commununications of ACM*, 40(3):56–58, 1997.

[158] Paul Resnick, Neophytos Iacovou, Mitesh Suchak, Peter Bergstrom, and John Riedl. Grouplens: an open architecture for collaborative filtering of netnews. In *Proceedings of the 1994 ACM conference on Computer Supported Cooperative Work*, CSCW '94, pages 175–186. ACM 1994

[159] Francesco Ricci, Lior Rokach, and Bracha Shapira. Introduction to recommender systems handbook. In *Recommender Systems Handbook*, pages 1–35. Springer, 2011.

[160] Elaine Rich. User modeling via stereotypes. *Cognitive science*, 3(4):329–354, 1979.

[161] Elaine Rich. Users are individuals: individualizing user models. *International journal of man-machine studies*, 18(3):199–214, 1983.

[162] Giuseppe Rizzo and Raphaël Troncy. NERD: Evaluating named entity recognition tools in the web of data. In *Proceedings of Workshop on Web Scale Knowledge Extraction*, ISWC'11, 2011.

[163] Daniel M. Romero, Brendan Meeder, and Jon Kleinberg. Differences in the mechanics of information diffusion across topics: Idioms, political hashtags, and complex contagion on twitter. In *Proceedings of the 20th International Conference on World Wide Web*, WWW '11, page 695–704. ACM, 2011.

[164] Matthew Rowe and Fabio Ciravegna. Getting to me: Exporting semantic social network from facebook. In *Proceedings of the 1st Workshop on Social Data on the Web*, 2008.

[165] Matthew Rowe and Milan Stankovic. Aligning tweets with events: Automation via semantics. *Semantic Web*, 3(2):115–130, 2012.

[166] Matthew Rowe, Milan Stankovic, and Philippe Laublet. Mapping tweets to conference talks: A goldmine for semantics. In *Proceedings of the 2nd Workshop on Social Data on the Web*, volume 664. CEUR-WS.org, 2010.

[167] Hassan Saif, Yulan He, and Harith Alani. Semantic sentiment analysis of twitter. In *Proceedings of the 11th International Conference on The Semantic Web - Volume Part I*, ISWC '12, pages 508–524. Springer-Verlag, 2012.

[168] Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. Earthquake shakes twitter users: real-time event detection by social sensors. In *Proceedings of the 19th International Conference on World wide web*, WWW '10, pages 851–860. ACM, 2010.

[169] Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. Earthquake shakes twitter users: real-time event detection by social sensors. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pages 851–860. ACM, 2010.

[170] Jagan Sankaranarayanan, Hanan Samet, Benjamin E. Teitler, Michael D. Lieberman, and Jon Sperling. Twitterstand: news in tweets. In *Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, GIS '09, pages 42–51. ACM, 2009.

[171] Badrul Sarwar, George Karypis, Joseph Konstan, and John Riedl. Item-based collaborative filtering recommendation algorithms. In *Proceedings of the 10th International Conference on World Wide Web*, WWW '01, pages 285–295. ACM, 2001.

[172] Sebastian Schaffert. Ikewiki: A semantic wiki for collaborative knowledge management. In *Proceedings of the 15th IEEE International Workshops on enabling Technologies: Infrastructure for Collaborative Enterprises*, WETICE '06, pages 388–396. IEEE, 2006.

[173] Andrew I. Schein, Alexandrin Popescul, Lyle H. Ungar, and David M. Pennock. Methods and metrics for cold-start recommendations. In *Proceedings of the 25th International ACM SIGIR conference on Research and Development in Information Retrieval*, SIGIR '02, pages 253–260. ACM, 2002.

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

[175] Ilaria Torre. Adaptive systems in the era of the semantic and social web, a survey. *User Modeling and User-Adapted Interaction*, 19(5):433–486, 2009.

[176] Andranik Tumasjan, Timm O. Sprenger, Philipp G. Sandner, and Isabell M. Welpe. Predicting elections with twitter: What 140 characters reveal about political sentiment. In *Proceedings of the 4th International Conference on Weblogs and Social Media*, ICWSM '10. The AAAI Press, 2010.

[177] Michal Tvarožek. Personalized navigation in the semantic web. In *Adaptive Hypermedia and Adaptive Web-Based Systems*, pages 467–471. Springer, 2006.

[178] Twitter. http://www.twitter.com.

[179] W3C Working Group. W3C resource description framework. Technical report, World Wide Web Consortium (W3C), October 1999.

[180] Wenbo Wang, Lu Chen, Krishnaprasad Thirunarayan, and Amit P. Sheth. Harnessing twitter "big data" for automatic emotion identification. In *Proceedings of the 2012 ASE/IEEE International Conference on Social Computing and 2012 ASE/IEEE International Conference on Privacy, Security, Risk and Trust*, SOCIALCOM-PASSAT '12, pages 587–592. IEEE 2012.

[181] Jianshu Weng and Bu-Sung Lee. Event Detection in Twitter. In *Proceedings of the 5th International AAAI Conference on Weblogs and Social Media*, ICWSM '11. The AAAI Press.

[182] Jianshu Weng, Ee-Peng Lim, Jing Jiang, and Qi He. Twitterrank: finding topic-sensitive influential twitterers. In *Proceedings of the Third International Conference on Web Search and Web Data Mining*, WSDM '10, pages 261–270. ACM, 2010.

[183] Jim Wissner and Nova Spivack. Case study: Twine. *W3C, Semantic Web Use Cases and Case Studies*, 2009.

[184] Sarita Yardi and Danah Boyd. Tweeting from the town square: Measuring geographic local networks. *Proceedings of the 4th International Conference on Weblogs and Social Media*, ICWSM '10. The AAAI Press, 2010.

[185] Shaozhi Ye and Felix Wu. Measuring message propagation and social influence on twitter.com. In *Proceedings of the Second International Conference on Social Informatics*, pages 216–231. Springer-Verlag, 2010.

[186] Youtube. http://www.youtube.com.

[187] Louis Yu, Sitaram Asur, and Bernardo A. Huberman. What trends in chinese social media. *CoRR*, abs/1107.3522, 2011.

[188] Eva Zangerle, Wolfgang Gassler, and Günther Specht. Recommending#-tags in twitter. In *Proceedings of the Workshop on Semantic Adaptive Social Web*, SASWEB '11, volume 730, pages 67–78. CEUR.org, 2011.

[189] Hua-Ping Zhang, Qun Liu, Hongkui Yu, Xueqi Cheng, and Shuo Bai. Chinese named entity recognition using role model. *Computational Linguistics and Chinese Language Processing*, volume 9, number 2, page 29–60. 2003.

List of Figures

3.1 Temporal constraint for microblogging-based user modeling 34
3.2 Impact of parameter d and timestamp $time_c$ on the weights of topic. 35
3.3 The architecture of GeniUS Library 38

4.1 Generic architecture for linking tweets with news articles and constructing user profiles. 50
4.2 Example of linking tweets with news articles and constructing user profiles. 52
4.3 Number of tweets per user $u \in U_u$ as well as the number of interactions (re-tweeting or reply activities) with Twitter accounts maintained by mainstream news media. 56
4.4 Precision of different strategies for relating tweets with news articles. 57
4.5 Number of tweets per user according to the different strategies related to news articles. 58
4.6 Comparison between tweet-based and news-based user modeling strategies for creating (a) entity-based profiles and (b) category-based profiles. 60
4.7 Comparison between tweet-based and news-based user modeling strategies with respect to (a) the variety of facet types available in the user profiles (example facet types: person, event, location, product) and (b) number of distinct hash tags and entities per profile. . . 61
4.8 Number of distinct emotion facets for individual users 69
4.9 The average entropy of emotion-based user profiles for different types of emotions 70

5.1 Number of distinct access clients for individual users 80
5.2 Comparison of writing style for individual users 82
5.3 Semantic analysis for individual users 85
5.4 The ration of positive posts on two microblogging services 87
5.5 Weekend-weekday ratio per user 89
5.6 Comparison of topic drift. 89

5.7 Overview of reposting behavior for individual users. 93
5.8 How fast do users repost? 94
5.9 Comparison of the interest focus 95
5.10 Comparison of the syntactic characteristics 95
5.11 Sentiment analysis of the reposts 96

6.1 Comparison between different user modeling strategies with tweet-only-based or news-based enrichment. 105
6.2 Temporal evolution of user profiles: average d_1-distance of current individual user profiles with corresponding profiles in the past. . . . 107
6.3 Temporal patterns: comparison between weekend and weekday profiles by means of d_1-distance ((a)-(c): category-based profiles). . . . 108
6.4 Results of news recommendation experiment. 111
6.5 Combining trend and user modeling to support trend-aware personalization 117
6.6 Standard deviation of different types of entities (cf. Equation 6.11). 118
6.7 Top entities for a given week. 119
6.8 Top trends for a given week. 119
6.9 Comparison of different weighting functions 122
6.10 Comparison of different type of profiles ($T_{t-TF\times IDF@500}(I_r)$, $P(u)$ and $M(I_r,u)$) 122
6.11 Comparison of different strategies for combining user and trend profiles 123
6.12 Relative occurrence frequencies of entities related to the Egyptian revolution. 126
6.13 User adoption: number of new users per day who become interested in the Egyptian revolution. 128
6.14 Daily activities of users who are interested in the Egyptian revolution. 128
6.15 Standard Deviation of Timestamps of Related Tweets Posted by Each User 129
6.16 Comparison of user modeling strategies for supporting personalization. 133
6.17 Relation between size of profiles and quality of profiles for supporting personalization. 134
6.18 Number of tweets, DBpedia entities and entity types per user 137
6.19 Comparison of different strategies for user profile construction . . . 138
6.20 Results of recommendation experiment 141

List of Tables

3.1 Design space of Twitter-based user modeling strategies. 28
3.2 Example microposts of a user 30
3.3 Topics of interests extracted from the example microposts in Table 3.2 30
3.4 Comparison of two design alternatives for semantic enrichment of microposts: (i) micropost-only-based, (ii) exploitation of external resources. 33
3.5 Example entity-based user profiles constructed with different strategies. 36
3.6 Hypotheses on the impact of different dimensions on the characteristics and quality of user profiles. 44

4.1 Conventional markers used to label emotion in microposts 64
4.2 Summary for classification of emotions. 66
4.3 Detailed results for classification of emotion with WordNet-based feature. 67
4.4 Emotion expressed in overall posts. 68
4.5 Entropy of example user profiles. 70
4.6 Overview on research questions investigated in this chapter. 71

5.1 Overview of datasets for Sina Weibo and Twitter. 79
5.2 Fraction of posts for different categories of clients 80
5.3 Comparison of syntactic content analysis 82
5.4 Impact of the access behavior on the syntactic characteristics of microposts. 83
5.5 Semantic analysis overall and impact of access behavior on the semantics. 85
5.6 Sentiment expressed in overall posts and posts that mention certain types of topics 86
5.7 Ratio between weekend posts and weekday posts 88
5.8 Hofstede's cultural index for China and United States 90
5.9 Some findings and their correlation with cultural dimensions. 98

6.1 Design dimensions and design alternatives that are evaluated in Section 6.2. 104
6.2 The average number of relevant items and candidate items for different recommendations. 140
6.3 Overview on research questions investigated in this chapter. 144

Summary

User Modeling and Personalization in the Microblogging Sphere

Microblogging has become a popular mechanism for people to publish, share, and propagate information on the Web. The massive amount of digital traces that people have left in the microblogging sphere, creates new possibilities and poses challenges for user modeling and personalization. How can microblogging activities be exploited to infer individual users' interests? How can semantically meaningful user profiles be constructed to support different applications? Does the users' microblogging behavior vary between different cultural groups? What is the impact of different user modeling strategies on the characteristics of user profiles and the performance of personalized recommendations?

In this thesis, we answer the research questions above and introduce a generic framework that provides a variety of user modeling strategies for inferring individual users' interests from microblogging streams. We propose and evaluate techniques that allow for exploiting external resources to enrich the semantics of short microblogging messages. We explore different approaches for deducing and modeling topics of interests based on enriched microblogging data. Furthermore, we investigate various weighting schemes for constructing user profiles and incorporate temporal constraints into the user modeling process.

With flexible design choices, the user modeling framework allows for constructing user profiles which can be consumed in different applications. We apply our user modeling framework to analyze user behavior across cultural groups on microblogging platforms. By exploiting different user characteristics, we unveil key differences in users' microblogging behavior between Chinese and American users. Finally, we analyze and evaluate different user modeling strategies in the context of various personalized recommender systems. The results of our analyses show that the characteristics of user profiles are significantly influenced by different design alternatives. In a set of experiments we reveal that the semantic enrichment of microposts and the consideration of temporal patterns improve the performance of

recommender systems. We also prove that the incorporation of public trends and domain specific knowledge into the user modeling process improves the quality of personalized recommendations.

Samenvatting

User Modeling and Personalization in the Microblogging Sphere

Microblogging is een populair mechanisme dat mensen hanteren om informatie op het Web te publiceren, delen en verspreiden. De grote hoeveelheid aan digitale sporen die mensen achterlaten in de ruimte van microblogs, leidt tot nieuwe mogelijkheden en uitdagingen voor het modelleren van gebruikers en personalisatie. Hoe kunnen microblog-activiteiten worden benut om de voorkeuren van individuele gebruikers vast te stellen? Hoe kunnen betekenisvolle gebruikersprofielen worden geconstrueerd voor verschillende toepassingen? Verschilt het gedrag van microblog-gebruikers over verschillende culturele groepen? Wat is het effect van verschillende strategieën voor gebruikersmodellering op de karakteristieken van gebruikersprofielen en de werking van systemen voor gepersonaliseerde aanbevelingen?

In dit proefschrift beantwoorden we deze onderzoeksvragen en introduceren we een algemeen raamwerk met een verscheidenheid aan strategieën voor gebruikersmodellering om de interesses van individuele gebruikers af te leiden uit stromen van microblogs. We presenteren en evalueren technieken die toestaan om externe bronnen te gebruiken om de semantiek van korte microblog-boodschappen te verrijken. We verkennen verschillende methoden voor het bepalen en modelleren van onderwerpen van interesse gebaseerd op verrijkte microblog-data. Verder onderzoeken we verschillende wegingsschema's voor het construeren van gebruikersprofielen en betrekken van tijd-constraints in het proces van gebruikersmodellering.

Door flexibele ontwerpkeuzes maakt het raamwerk voor gebruikersmodellering het mogelijk om gebruikersprofielen te construeren voor verschillende toepassingen. We passen ons raamwerk voor gebruikersmodellering toe om gebruikersgedrag te analyseren binnen verschillende culturele groepen in platforms voor microblogging. Door verschillende karakteristieken van gebruikers te benutten zijn we in staat belangrijke verschillen vast te stellen in het gedrag tussen Chinese en Amerikaanse gebruikers. Tenslotte analyseren en evalueren we verschillende strategieën voor gebruikersmodellering in de context van systemen voor gepersonaliseerde aan-

bevelingen. De resultaten van onze analyses tonen aan dat de karakteristieken van gebruikersprofielen significant worden beïnvloed door verschillende ontwerpkeuzen. Met een reeks van experimenten laten we zien dat de semantische verrijking van microposts en het beschouwen van temporele patronen de werking verbeteren van systemen voor aanbevelingen. We laten ook zien dat het gebruik van publieke trends en domeinspecifieke kennis in het proces van gebruikersmodellering de kwaliteit van persoonlijke aanbevelingen verbetert.

Curriculum Vitae

Qi Gao was born in Zhejiang, China on July 28, 1982. He obtained his B.E. degree in 2006 at Tongji University, Shanghai, China. Between 2006 and 2009, he was a research assistant at Tongji University.

From Sep. 2009 to Oct. 2013, he was a Ph.D. student in the Web Information Systems group at Delft University of Technology, the Netherlands, supervised by Prof. Geert-Jan Hoube. During his Ph.D., he conducted research on user modeling and personalization based on large amount of microblogging data. He has won the best paper award at UMAP2011 and the James Chen best student paper award at UMAP2012.

Publications

- Fabian Abel, Qi Gao, Geert-Jan Houben, Ke Tao. *Twitter-Based User Modeling for News Recommendations*. In Proceedings of International Joint Conference on Artificial Intelligence (IJCAI), Beijing, Shanghai, 2013.
- Qi Gao, Fabian Abel, Geert-Jan Houben, Yong Yu. *A Comparative Study of Users Microblogging Behavior on Sina Weibo and Twitter*. In Proceedings of International Conference on User Modeling and Personalization (UMAP), Montral, Canada, 2012.
- Qi Gao, Fabian Abel, Geert-Jan Houben, Yong Yu. *Information Propagation Cultures on Sina Weibo and Twitter*. In Proceedings of ACM Conference on Web Science (WebSci), Evanston, USA, 2012.
- Qi Gao, Fabian Abel and Geert-Jan Houben. *GeniUS : Generic User Modeling Library for the Social Semantic Web*. In Joint International Semantic Technology Conference (JIST), Hangzhou, China, 2011.
- Fabian Abel, Qi Gao, Geert-Jan Houben and Ke Tao. *Analyzing User Modeling on Twitter for Personalized News Recommendations*. In Proceedings of International Conference on User Modeling, Adaptation and Personalization (UMAP), Girona, Spain, 2011.

- Fabian Abel, Samur Aurojo, Qi Gao and Geert-Jan Houben. *Analyzing Cross-System User Modeling on the Social Web*. In Proceedings of the Eleventh International Conference on Web Engineering (ICWE), Paphos, Cyprus, 2011.
- Qi Gao, Fabian Abel, Geert-Jan Houben and Ke Tao. *Interweaving Trend and User Modeling for Personalized News Recommendation*. In Proceedings of IEEE/WIC/ACM International Conference on Web Intelligence, Lyon, France, 2011.
- Fabian Abel, Qi Gao, Geert-Jan Houben and Ke Tao. *Analyzing Temporal Dynamics in Twitter Profiles for Personalized Recommendations in the Social Web*. In Proceedings of International Conference on Web Science(WebSci), Koblenz, Germany, 2011.
- Fabian Abel, Qi Gao, Geert-Jan Houben and Ke Tao. *Semantic Enrichment of Twitter Posts for User Profile Construction on the Social Web*. In Proceedings of Extended Semantic Web Conference (ESWC), Heraklion, Greece, 2011.
- Ke Tao, Fabian Abel, Qi Gao and Geert-Jan Houben. *TUMS : Twitter-based User Modeling Service*. In Proceedings of International Workshop on User Profile Data on the Social Semantic Web (UWeb) at ESWC2011, Heraklion, Greece, 2011.
- Qi Gao. *Towards Trust in Web Content Using Semantic Web Technologies*. In Proceedings of 8th Extended Semantic Web Conference(ESWC), Heraklion, Greece, 2010.
- Qi Gao and Geert-Jan Houben. *A Framework for Trust Establishment and Assessment on the Web of Data*. In Proceedings of the 19th international conference on World Wide Web(WWW), Raleigh, North Carolina, 2010.
- Samur Araújo, Qi Gao, Erwin Leonardi and Geert-Jan Houben. *Carbon : Domain-Independent Automatic Web Form Filling*. In Proceedings of International Conference on Web Engineering(ICWE), Vienna, Austria, 2010.
- Qi Gao, Junwei Yan and Min Liu. *A Semantic Approach to Recommendation System Based on User Ontology and Spreading Activation Model*. In Proceedings of NPC Workshops, Shanghai, China, 2009.
- Min Liu, Qi Gao, Weiming Shen, Qi Hao and Junwei Yan. *A Semantic-augmented Multi-level Matching Model of Web Services*. Service Oriented Computing and Applications 3(3), 2009.
- Min Liu, Weiming Shen, Qi Hao, Junwei Yan and Qi Gao. *A Multi-level Matching Framework for Semantic Web Services in Collaborative Design*. In Proceedings of 12th International Conference on CSCW in Design(CSCWD), Xian, China, 2008.