User Expertise Modelling
Using Social Network Data

A thesis submitted to the
University of Dublin, Trinity College
in fulfilment of the requirements of the degree of
Doctor of Philosophy

Yu Xu
Knowledge and Data Engineering Group (KDEG)
School of Computer Science and Statistics
Trinity College, Dublin
Ireland

Supervised by Prof. Séamus Lawless

2018
Declaration

I, the undersigned, declare that this work has not been previously submitted as an exercise for a degree at this or any other University, and that, unless otherwise stated, it is entirely my own work.

Signature ___________________________ Date __________________

Yu Xu
Permission to lend or copy

I, the undersigned, agree that the Trinity College Library may lend or copy this thesis upon request.

Signature ___________________________  Date ________________________

Yu Xu
Acknowledgements

I would like to express my appreciation to numerous people who have helped me in ways large and small to complete my work over the last number of years.

Firstly, I would like to express my sincere gratitude to my advisor Professor Séamus Lawless for all his support and guidance personally and professionally throughout this Ph.D. I could not have done it without him.

I would like to thank all my colleagues in KDEG and ADAPT at Trinity College Dublin for their help both personally and academically, including all the Irish language and culture lessons! They truly made me feel like I am part of the Irish society.

I would also like to thank my family for all their support, patience, constant encouragement all along the way.

Finally, I would like to express my gratitude to Science Foundation Ireland for funding the research detailed in this thesis.
Abstract

The ability to understand the expertise of online users is a key component for delivering effective information services such as talent seeking and user recommendation. However, users are often unwilling to make the effort to explicitly provide this information. Data extracted from social networking sites can be a valuable resource for inferring this expertise. But research on user expertise modelling in social networking sites is still immature and faces a number of critical challenges:

1) Modelling the expertise of cold start users;
2) The effective utilisation of a variety of user data for user expertise modelling;
3) The effective application of the modelled user expertise information.

The contribution of this thesis is concerned with addressing aspects of these three challenges. Firstly, this research proposes to use the static social media profile of a cold start user to model their language information. This is mainly based on the intuition that a user's experiences could imply what languages they know. A language and social relation-based factor graph model is proposed which exploits the dependency relations between languages as well as social relations between profiles to better model this problem. Secondly, this research focuses on the popular micro-blogging site Twitter and aims to model the general topics of expertise a user has knowledge of, based upon their various user data. Specifically, sentiment analysis is first used to assess the importance of each tweet, the primary data of a Twitter user, in user expertise modelling. Then, based on discriminative learning, this research proposes a model that infers a user's expertise under the collective but discriminative influence of various data related to the user. The proposed model exploits the dependency relations between expertise topics and consistency relations between different types of user data in order to better infer the user’s topical expertise. Thirdly, this research studies the application of the proposed user expertise modelling methods in the scenario of a community question answering site. It aims to model the expertise information of more users in the platform through the adoption of the proposed methods and help to more effectively find potential answerers to newly posted questions, which offers an effective way to improve the question answering service.
# Table of Contents

Declaration ............................................................................................................................ ii
Permission to lend or copy ....................................................................................................... iii
Acknowledgements ................................................................................................................ iv
Abstract ................................................................................................................................... v
Table of Contents .................................................................................................................. vi
List of Figures ......................................................................................................................... x
List of Tables .......................................................................................................................... xii
Abbreviations ......................................................................................................................... xiii

1. Introduction.......................................................................................................................... 1
   1.1. Motivation .................................................................................................................... 1
   1.2. Research Question and Challenges ........................................................................... 2
   1.3. Research Objectives ................................................................................................... 5
   1.4. Thesis Contribution .................................................................................................... 6
   1.5. Thesis Overview ......................................................................................................... 10

2. State of the Art .................................................................................................................... 12
   2.1. User Expertise Modelling in Social Networking Sites ................................................. 12
      2.1.1. User Expertise Modelling on Twitter ................................................................. 12
      2.1.1.1. User Influence Modelling ............................................................................... 12
      2.1.1.2. User Expertise Modelling on Latent Topics .................................................. 15
      2.1.1.3. User Expertise Modelling on Specific Topics ............................................... 17
      2.1.2. User Expertise Modelling in Other SNSs ........................................................... 21
      2.1.3. Section Summary ................................................................................................. 23
   2.2. Expertise Modelling Methods ..................................................................................... 24
      2.2.1. Generative Models ............................................................................................... 24
      2.2.1.1. Candidate Generation Models ........................................................................ 25
      2.2.1.2. Topic Generation Models ................................................................................ 27
      2.2.2. Discriminative Models ......................................................................................... 28
      2.2.2.1. A General Discriminative Learning Framework ............................................ 29
      2.2.2.2. Learning to Rank ............................................................................................. 30
2.2.2.3. Other models ................................................................. 31
2.2.3. Section Summary ............................................................. 32

2.3. Answerer Finding in CQA Sites ........................................... 33
  2.3.1. Problem Background .................................................... 33
  2.3.2. Answerer Finding Approaches ....................................... 34
     2.3.2.1. Language Model-based Approaches ............................ 34
     2.3.2.2. Topic Model-based Approaches ................................ 35
     2.3.2.3. Classification-based Approaches ............................... 37
     2.3.2.4. Other Approaches .................................................. 38
  2.3.3. Section Summary ....................................................... 39

2.4. Other Related Work ......................................................... 40
  2.4.1. User Language Information Inference ............................. 40
  2.4.2. Tweet Sentiment Analysis .......................................... 40

2.5. Chapter Summary ............................................................ 42

3. Language Information Inference Using Social Profiles .............. 44
  3.1. Introduction ........................................................................ 44
  3.2. Problem Description ........................................................ 46
     3.2.1. Background .............................................................. 47
     3.2.2. Challenges and Solutions ......................................... 47
     3.2.3. Problem Definition ................................................... 50
  3.3. Language and Social Relation-based Factor Graph Model ...... 50
     3.3.1. Model Definition ....................................................... 50
     3.3.2. Model Learning and Prediction ................................... 54
  3.4. Experiment and Results .................................................... 55
     3.4.1. Dataset Construction and Location Information Enhancement .... 55
     3.4.2. Baseline Methods .................................................... 56
     3.4.3. Evaluation Metrics ................................................... 57
     3.4.4. Performance Comparison and Analysis .......................... 58
3.4.5. Factor Contribution Analysis .................................................. 60

3.5. Chapter Summary ........................................................................... 60

4. Topical Expertise Inference on Twitter: Integrating Sentiment and Topic Relatedness ................................................................. 62
   4.1. Introduction .................................................................................. 62

   4.2. Sentiment-Weighted and Topic Relation-Regularized Learning for User Expertise Inference ......................................................... 64
   4.2.1. Problem Definition .................................................................. 64
   4.2.2. Feature Weighting with Tweet Sentiment Analysis .................. 65
   4.2.3. Expertise Inference with Topic Relation Regularization .......... 66
   4.2.4. Model Learning and Prediction ............................................... 68
   4.2.5. Construction of Topic Relation Structure ............................... 70

   4.3. Dataset Construction ................................................................. 70

   4.4. Experiment and Results .............................................................. 72
   4.4.1. Baseline Methods and Evaluation Metrics ................................. 72
   4.4.2. Parameter Configuration .......................................................... 73
   4.4.3. Performance Comparison and Analysis .................................... 75

   4.5. Chapter Summary ........................................................................ 78

5. Topical Expertise Inference on Twitter: Combining Multiple Types of User Activity ................................................................. 80
   5.1. Introduction .................................................................................. 80

   5.2. User Expertise Inference Using Multiple Types of Data on Twitter ........... 82
   5.2.1. Problem Definition ................................................................. 82
   5.2.2. Model Definition ..................................................................... 83

   5.3. Dataset Construction and Experiment Preparation ..................... 86

   5.4. Experiments and Results .............................................................. 89
   5.4.1. Evaluation Metrics .................................................................. 89
   5.4.2. Performance of Different Types of User Data ............................ 89
   5.4.3. Performance of Combinations of Different Types of User Data .... 95
5.5. Chapter Summary ................................................................................................................................. 97

   6.1. Introduction ........................................................................................................................................ 98

6.2. Answerer Finding on CQA Sites ..................................................................................................... 100
       6.2.1. User Answering History Based Approach .............................................................................. 100
       6.2.2. User Expertise Topic Based Approach ................................................................................. 101

6.3. Dataset Construction ...................................................................................................................... 102

6.4. Evaluation Method ........................................................................................................................ 106
       6.4.1. Ground Truth Answerers Construction ................................................................................. 106
       6.4.2. Evaluation Metric and Comparative Methods ..................................................................... 108

6.5. Experiments and Results ............................................................................................................... 110
       6.5.1. Performance of the User Answering History Based Approach ............................................. 115
       6.5.2. Effectiveness of the Specified User Expertise Topics ......................................................... 116
       6.5.3. Effectiveness of the Inferred User Expertise Topics ......................................................... 116

6.6. Chapter Summary ............................................................................................................................ 117

7. Conclusion .............................................................................................................................................. 119
   7.1. Research Objectives vs Research Achievements ............................................................................. 119
       7.1.1. Research Objectives Recap .................................................................................................. 119
       7.1.2. Evaluation of Research Achievement .................................................................................. 120
       7.1.2.1. Objective 1: Expertise Inference of Cold Start Users ...................................................... 120
       7.1.2.2. Objective 2: Expertise Inference of Active Users ............................................................ 121
       7.1.2.3. Objective 3: Application of the Inferred Expertise Information ..................................... 123

   7.2. Contributions ................................................................................................................................... 124

   7.3. Further Work ....................................................................................................................................... 126
       7.3.1. Short-Term Future Work ..................................................................................................... 126
       7.3.2. Long-Term Future Work ..................................................................................................... 128

   7.4. Final Remarks ..................................................................................................................................... 129

References .................................................................................................................................................. 130
List of Figures

Figure 1-1: A Framework of Expertise Discovery in SNSs and Its Application in CQA sites .......................................................... 7
Figure 3-1: Graphic representation of the LSR-FGM ......................................................... 51
Figure 4-1: An example of a Quora Profile ................................................................. 71
Figure 4-2: F1 performance of SVM-Sen and SVM with the different sentiment_ threshold .................................................................................. 74
Figure 4-3: Accuracy performance of SVM-Sen and SVM with the different sentiment_ threshold .............................................................................. 74
Figure 4-4: Performance of SeTRL-Co with different α .................................................. 75
Figure 4-5: Performance of SeTRL-Co with different β .................................................. 75
Figure 4-6: F1 performance of different methods on topics with different sizes of training data ......................................................................................................................... 77
Figure 4-7: F1 performance of different methods on the topics "Mobile Marketing", "Public Relations", "Machine Learning" .......................................................... 78
Figure 5-1: Distribution of the friend numbers of Twitter users in the dataset .......... 87
Figure 5-2: Distribution of the follower numbers of Twitter users in the dataset .... 87
Figure 5-3: Distribution of the list numbers of Twitter users in the dataset .......... 88
Figure 5-4: Performance of SVM with tweet data and latent topic features .......... 93
Figure 5-5: Performance of SVM with friend data and latent topic features .......... 93
Figure 5-6: Performance of SVM with follower data and latent topic features .......... 94
Figure 5-7: Performance of SVM with list data and latent topic features .......... 94
Figure 6-1: An example of a Quora user’s answer page .............................................. 103
Figure 6-2: An example of the page of the full answer to a question ...................... 103
Figure 6-3: An example of a question page with knowledge topics ....................... 105
Figure 6-4: Ranking accuracy of different answerer finding approaches for the topic “sales” .................................................................................................................. 110
Figure 6-5: Ranking accuracy of different answerer finding approaches for the topic “machine learning” . ................................................................. 111
Figure 6-6: Ranking accuracy of different answerer finding approaches for the topic “video games” ................................................................................ 111
Figure 6-7: Ranking accuracy of different answerer finding approaches for the topic “music” ........................................................................................................... 112
Figure 6-8: Ranking accuracy of different answerer finding approaches for the topic “recruiting” ........................................................................................................ 112
Figure 6-9: Ranking accuracy of different answerer finding approaches for the topic “photography” ................................................................................. 113
Figure 6-10: Ranking accuracy of different answerer finding approaches for the topic “philosophy” .................................................................................... 113
Figure 6-11: Ranking accuracy of different answerer finding approaches for the topic “islam” ......................................................................................................... 114
Figure 6-12: Ranking accuracy of different answerer finding approaches for the topic “startups” ......................................................................................... 114
Figure 6-13: Ranking accuracy of different answerer finding approaches for the topic “dating and relationships” ................................................. 115
List of Tables

Table 3-1: Definition of Notations ................................................................. 49
Table 3-2: The distribution of languages in the collected social profiles .......... 55
Table 3-3: An example of the confusion matrix for the prediction of user expertise on language $l$ ........................................................................................................ 57
Table 3-4 Performance of language inference with different methods on different languages (%) ............................................................................................................ 58
Table 3-5: Overall performance of LSR-RGM with different factors (%) .......... 60
Table 4-1: Definition of Notations ................................................................. 65
Table 4-2: Statistics of the dataset ..................................................................... 72
Table 4-3: Overall performance of expertise inference with different methods (%) ................................................................. 76
Table 5-1: Definition of Notations ................................................................. 83
Table 5-2: Performance of SVM method with each type of user data for user expertise inference (%) ............................................................................................................ 91
Table 5-3: Performance of TRL method with each type of user data and unigram features (%) ............................................................................................................ 92
Table 5-4: Performance of TRL method with each type of user data and user latent topic features (%) ............................................................................................................ 95
Table 5-5: Performance of different methods with the combination of the four types of user data (%) ............................................................................................................ 96
Table 5-6: Performance of D*TCom with various combinations of different types of user data (%) ............................................................................................................ 96
Table 6-1: Distribution of the number of questions collected per user .......... 104
Table 6-2: Statistics of the testing samples by expertise topics .................... 107
### Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Full Form</th>
</tr>
</thead>
<tbody>
<tr>
<td>SNS</td>
<td>Social Networking Site</td>
</tr>
<tr>
<td>CQA</td>
<td>Community Question Answering</td>
</tr>
<tr>
<td>IR</td>
<td>Information Retrieval</td>
</tr>
<tr>
<td>EF</td>
<td>Expert Finding</td>
</tr>
<tr>
<td>AF</td>
<td>Answerer Finding</td>
</tr>
<tr>
<td>L2R</td>
<td>Learning to rank</td>
</tr>
<tr>
<td>UAH</td>
<td>User Answering History</td>
</tr>
<tr>
<td>UET</td>
<td>User Expertise Topic</td>
</tr>
</tbody>
</table>
1. Introduction

1.1. Motivation

Finding individuals with a required set of expertise is a common task, such as when searching for a solution to a problem, attempting to perform a given task, or seeking a product recommendation. Situations which involve this task are commonplace. For example:

- In online employee recruitment, such as in LinkedIn, the *professional skills* of potential employees are an essential aspect for personnel resource managers to examine;

- Knowledge of what *languages* an online user comprehends is the key to enabling search engines to deliver multilingual search services, machine translation tools to identify optional target translation languages, and advertisers to serve targeted international advertisements;

- Knowing what topics an online user has knowledge of eases the matching between questions and answerers in a question-answering scenario.

Thus, the core challenge tackled by this research is to design user modelling approaches which can accurately infer a user’s expertise, such as the languages they speak, their professional skills, or even general topics of expertise.

A large proportion of the existing studies in this area have targeted the task of *expert finding* within an organization [Ma06a, Ba06a, Ba12]. In this scenario, they model the level of expertise of people with respect to different topics based on their documents within the organization such as project descriptions, program codes, emails, or the employee’s own files [e.g. Be00, Po01a, Re09, Vi00]. This knowledge can greatly benefit the effective identification of domain experts in the organization for carrying out given tasks or solving specific issues. In addition, a number of studies also explored platforms, such as academic networks [Zh07a, De08], blogs [Ka07, Na08] and question answering sites [Zh07b, Ju07], and try to help users find experts on specific topics. By doing this, different platforms aim to benefit users in different ways. For example, a platform could help researchers efficiently find collaborators in academic circles; another could help to locate potential answerers to questions in question answering sites.
In recent years, Social Networking Sites (SNSs) such as Facebook and Twitter have been integrated into people's daily lives. People share experiences, catch up on the activities of their friends or directly communicate with them. Such platforms greatly facilitate communication and information exchange between people, and offer a new way of making and maintaining friendships. However, as the prevalence of SNSs increases, the SNS is not only playing the role of a simple communication tool, but also becoming a comprehensive platform that integrates with various services and functionalities for facilitating different aspects of people's lives [Ki11], such as information search [Ho10a, Ef11], job hunting [El07, Su13] and product recommendation [Ch11, Ta13]. For some information services, discovering the expertise of SNS users is an important component. For instance, it is reported that 29% of companies match their employees to tasks using social media tools [Bu11]; SNS users often follow others in order to obtain professional knowledge [Ra10]; it is also common that people turn to their social networks when they have questions [Mo10, Ni12]. To enable services which can satisfy such information needs, obtaining expertise information for SNS users is key.

Expertise information is typically not explicitly provided by SNS users, so existing methods aimed at expertise discovery in SNSs primarily focus on implicit inference. The rich activities which users conduct on SNSs lead to the creation of an enormous amount of user-generated content and diverse associations among users and content, which provide valuable sources of information for user expertise inference. SNSs also have other unique advantages in expertise discovery. Firstly, the content and associations in SNSs are usually publicly available, such as most Twitter data and LinkedIn profiles, while personal documents within an organization, such as emails and local files are usually inaccessible or involve more complex privacy issues. Secondly, data in SNSs is dynamic in nature. Thus, the change or expansion of users' expertise can be captured over time. Finally, the worldwide popularity of SNSs has produced a huge living knowledge repository which is easily accessible to many people.

1.2. Research Question and Challenges

The question which this research aims to answer is:

*To what extent can a user’s content, actions and connections on social networking sites be exploited by novel user modelling approaches to infer their expertise?*
To assess whether the user expertise inferred by the proposed approaches is of benefit to information services, a further question will be addressed:

*Can the inferred user expertise be used to facilitate services such as answerer finding on a community question answering site?*

A number of attempts [Pa11, Wa12, Gu13, Po13, Li15] have been made in recent years to try to answer these, or similar questions, however such techniques for inferring a user's expertise information on social networking sites still face a number of critical challenges:

1. **Cold Start Users**

   Prior research has focused on the use of abundant user-generated data on SNSs for expertise inference, for example, messages authored or republished by a user, biographical information of a user, information about a user's friends, etc. [Pa11]. However, recent studies [Ba15] revealed that over 50% of content in popular SNSs, like Facebook and Twitter is produced by less than 10% of users, which indicates that a large proportion of users are relatively inactive and publish very little content. The ability to infer expertise information for such cold start users with limited information has received little attention, for example, a user with a static profile which they filled in when registering, but with no additional posts or tweets. As a result, it is a significant challenge to investigate what types of expertise information can be inferred in this scenario.

2. **Expertise Inference of Non-Cold Start Users**

   In the literature, there are two primary groups of approaches that try to model online user’s expertise information by exploiting their various activities on SNSs. The first group of approaches seeks to learn latent factors that could be descriptive of the user. In general, topic modelling approaches [Ra10, Wa12, Ho10b] are employed on a user’s social media content, e.g. tweets, to reveal the user’s expertise on different latent topics. In this case, user expertise is often represented as a probability distribution over latent topics which are also distributions over features in a lower dimensional space. The essential challenge of this group of approaches is that the latent topics are difficult to interpret, i.e. not directly understandable to humans.

   The second group of approaches focuses on inferring the user’s expertise on specific expertise topics, e.g. “machine learning”, by utilising their various social activities. In this group of approaches, there are two major threads of methods that try to reach the goal
of inferring the user’s specific expertise topics. The first thread of methods simplifies it as an Information Retrieval (IR) problem [We16, Ch14], i.e. given an expertise topic (query terms) and a list of users (documents built from the content, connections and actions related to each user), the objective is to compute the degree of relevance between the expertise need and each user. This thread of methods can obtain a user ranking with respect to an expertise need, with the expectation being that more knowledgeable people on a certain topic are ranked highest. However, this thread of methods suffers from the problem of clearly discriminating experts from other users. Besides, in social media environments, users can publish content on any topic, so the frequent mentions of a topic in a user's authored content does not necessarily indicate that this user has a good level of knowledge of that topic. More factors should be considered in the process of evaluating the level of expertise of a user, such as comments from other users on the current user-authored content and how many/what people republished the content. The second thread of approaches is Supervised Machine Learning (SML) based algorithms. This thread of methods often suffers from a shortage of labeled data, because most SNSs do not have an expertise attribute in user profiles and SNS users also typically do not want to go to the effort of authoring and maintaining this information. The existing studies that apply this method mainly rely on manual labeling for obtaining training data [Po13], which is always expensive and time-consuming. Another challenge facing this thread of methods is the effective selection and utilisation of features from available user data. Social content is noisy. To effectively identify discriminative features for each expertise topic is always challenging. In addition, previous research [Wa12] studied the usefulness of different types of user-generated data for inferring topical expertise of Twitter users and revealed that different types of user data differ in their ability to help expertise inference. These findings suggest that discriminatively utilising different types of user data will play an important role in inferring a user's expertise information.

(3) Applications of User Expertise Information

As introduced earlier, the effective modelling of SNS user expertise can serve many information services. This research focuses on one specific application scenario: Answerer Finding (AF) in Community Question Answering (CQA) sites, and studies the application of the inferred user expertise by our proposed modelling approaches in this area.
A CQA site allows users to answer questions asked by other users, e.g. Quora\(^1\) and Yahoo! Answers\(^2\). To offer a better CQA service and attract more users to these sites, one of the most critical factors is to help the question askers solve their questions efficiently. In other words, it is required to efficiently find the users who are capable of answering the asked questions. The most commonly used way of addressing this problem is to build the user’s expertise model by exploiting their answering history in CQA sites, and then match them with new questions [Zh09, Li10a, Li10b]. However, as revealed in a previous study [Sr15] the answering rate of new questions on CQA sites is still very low, i.e. only about 20\% of questions were answered within two days on Yahoo! Answers. One of the important reasons for this, is that on CQA sites a large proportion of users have answered very few, or never answered any, questions, i.e. are cold start users. Therefore, the expertise of these users cannot be modelled and potential answerers of questions are likely being ignored.

On the other hand, a user’s expertise can also be inferred from their social activities as discussed above. Also, there are mechanisms that allow us to connect the same user’s CQA account to their SNS accounts, for example, users can provide their Twitter and Facebook accounts on their Quora profiles; many users directly use their SNS accounts (e.g. Twitter account) to sign into a CQA site (i.e. social login\(^3\)) to avoid the hassle of filling in registration forms and keeping additional login information. Based on these facts, it naturally brings us the idea of modelling the expertise of cold start users on CQA sites from their social content and enabling them to be considered in answerer finding. By doing so, it could help to expand the search scope in answerer finding and increase the chance of finding more potential answerers to a question. However, the feasibility of applying the inferred user expertise information from SNSs to answerer finding on CQA sites remains untested. This calls for studies to investigate the usefulness of the inferred expertise information in the application scenario of answerer finding on CAQ sites.

1.3. Research Objectives

To overcome the three main challenges described above, and to answer the proposed research questions, the following objectives will be addressed:

---

\(^{1}\) [https://www.quora.com](https://www.quora.com)

\(^{2}\) [https://answers.yahoo.com/](https://answers.yahoo.com/)

A large proportion of the population of SNSs are cold-start users. In order to make the best use of the enormous social resources available, it is essential to investigate the possibility of discovering expertise from limited user-generated content, connections and actions. If this is possible, algorithms should be devised and evaluated for effective user expertise inference.

There is a diversity of expertise topics a user can have knowledge of on the Web, so it is unrealistic to human-label training data for each of the topics. This demands schemes of automatically constructing large-scale user-topic training data by exploiting available Web resources. Social content is noisy and diverse, and different types of user data contribute to the expertise inference discriminately. Thus, advanced user modelling approaches that can effectively exploit various users’ social data to infer their expertise information need to be developed and evaluated.

This PhD research takes answerer finding on CQA sites as the target application scenario. To offer a better question answering service, it is important to understand the expertise background of users on CQA sites and efficiently locate answerers to the newly asked questions. In social networking sites, the proposed user expertise modelling approaches make it possible to infer the user’s expertise information from their various social activities. In this context, new experimental methods need to be designed to investigate the usefulness of the inferred user expertise in the application scenario of answerer finding on CQA sites and this usefulness determined through evaluation.

1.4. Thesis Contribution
This PhD research includes a major contribution and a minor contribution.

**Major contribution:** the design of three user modelling approaches that infer a user’s expertise by exploiting their various social content on SNSs: (1) a Language and Social Relation-based Factor Graph Model (LSR-FGM) that aims to exploit the user’s static social profiles to infer information about the languages they speak; (2) a Sentiment-weighted and Topic Relation-regularized Learning (SeTRL) model that aims to exploit the user’s posted social content to infer their topical expertise and (3) a multi-Data and Topic relatedness Combined (DnTCom) learning model that aims to discriminately exploit multiple types of user data on SNSs to better infer the user’s topical expertise.

**Minor contribution:** the design of an experimental method that investigates the usefulness of the inferred expertise information about the user.
Fig. 1-1 illustrates the two aspects of the research: user expertise modelling approaches and the application of user expertise in answerer finding on CQA sites. Step 3 is the focus of the PhD research, i.e. the design of the three user expertise modelling approaches; Step 4 is the generated user expertise models using the proposed approaches; user expertise models are passed on to Step 5 to represent the expertise of cold start users in a CQA site if they are linked to their SNSs.

The major contribution of this PhD research considers two scenarios for user expertise modelling in SNSs: cold start users and non-cold start users.

In the cold start scenario, this research aims to infer possible expertise information about the user by exploiting the limited information that is available. In particular, it focuses on the exploitation of static structured user profiles. In many SNSs, such as Facebook and LinkedIn, users are asked to provide some basic personal information such as education.
and work experience when registering. These static profiles provide first-hand information about a new user to the SNS and connected applications. Intuitively, there is a chance that this information may implicitly suggest some expertise information about a user, e.g. what languages they speak, what professional skills they may have, etc. Thus, this research studies the use of static profiles for inferring certain types of expertise information about the user, with a focus on language expertise. An advanced modelling approach, called Language and Social Relation-based Factor Graph Model (LSR-FGM), is proposed to address the problem. It explores external resources to bring in more evidential signals, and exploits the dependency relations between languages as well as social relations between profiles in modelling the problem. Experiments are conducted on a large-scale dataset. The results demonstrate the success of the proposed modelling approach in language inference and show that it outperforms several alternative methods.

For non-cold start users, this PhD research focuses on the effective use of various social activities in user expertise modelling, with the aim of inferring the user’s topical expertise. Firstly, to address the challenge of the lack of labelled data in this field, this research proposes an approach to automatically annotate the expertise topics of SNS users through the popular CQA site Quora. This is because a large proportion of Quora users explicitly provide both their expertise topics and SNS accounts, e.g. Twitter and Facebook, on their profiles. This inspires us first to harvest a large volume of Quora profiles, from which we can obtain both expertise topics and SNS accounts for each individual. Twitter is the most commonly targeted SNS for research, due to the public availability of its user data. This research builds a dataset with over 10,000 Twitter users who have expertise in 149 different topics.

This research also looks at the problem of using the user’s posted content in SNSs to infer their expertise on a variety of topics. To overcome the effect of the noise in social content in modelling this problem, a Sentiment-weighted and Topic Relation-regularized Learning (SeTRL) model is proposed. The SeTRL model uses the sentiment intensity contained in a user’s posted content to evaluate its importance in inferring the user’s expertise. It also incorporates the relatedness between expertise topics in the process of user expertise inference. Experiments conducted on our Twitter dataset demonstrate the success of our proposed modelling approach in user expertise inference using the user’s tweets, and show that it outperforms the state-of-the-art modelling approaches.
Although shown to be effective in inferring a user’s expertise information by exploiting certain types of user data on SNSs (e.g. the user’s posted content, friends of the user, etc.), a fact that is often ignored is that many users of SNSs do not have common types of data. For example, on Twitter it is reported that about 44% of all registered users have never posted a tweet\(^4\). To overcome this issue in user expertise inference, this research proposes a multi-Data and Topic relatedness Combined (D\(^\text{mTCom}\)) learning model that infers a user’s topical expertise by utilising multiple types of user data on a SNS. The D\(^\text{mTCom}\) takes multiple types of user-related data from a SNS as input and considers their inference consistency in the process of learning. It aims to deliver accurate and effective inference results, even in cases where some types of data are missing for a user, e.g. the user has yet to post any tweets. Experiments conducted on the generated Twitter dataset show that the proposed D\(^\text{mTCom}\) learning model outperforms several baseline approaches and outperforms approaches which use only a single type of user data for inference.

A minor contribution of this PhD research is to study the usefulness of the proposed user expertise modelling approaches in a real-world application scenario. This research focuses on the application scenario of answerer finding on CQA sites.

As introduced earlier, many Quora users provide their SNS accounts on their profiles, which enables us to access the Quora user’s social content, e.g. public LinkedIn profiles and social activities conducted on Twitter. However, a large proportion of users on Quora have little question answering history, which makes it difficult to model their expertise information. Therefore, this research proposes to infer the cold start Quora user’s expertise from their social content by employing our proposed modelling approaches, and match them with new questions they are capable of answering. Experiments are conducted on a Quora dataset with 591,654 answered questions and 24,285 Quora users and results show that the inferred expertise information about the users can help to find desired answerers of new questions by only relying on their social activities from Twitter. Experiments also compared the performance of the answerer finding approach using the inferred expertise information with that of the user answering history based approach, and results show that the former achieves performance close to that of the latter.

The following four papers were published in relation to research described in this PhD thesis.


This paper defines the problem of user language inference using social profiles, and describes the proposed LSR-FGM as well as corresponding experimental results.


This paper describes the construction of dataset of Twitter users versus expertise topics, the proposed SeTRL model and related experimental results.


These two papers investigate the use of different types of user data on Twitter in user expertise inference and describe in detail the proposed D^2TCom learning model that infers the user’s topical expertise by combining multiple types of user data.

1.5. Thesis Overview

The remainder of this PhD thesis is structured as follows:

Chapter 2 presents the state of the art in user expertise modelling. It first reviews research that exploits SNS information to model a user's expertise. Then, alternative methodologies for user expertise modelling are reviewed regardless of the application platforms. Finally, this chapter reviews works related to the application of answerer finding on CQA sites and some other techniques applied in this PhD research.
Chapter 3 formally defines the problem of language information inference from the user’s static social profiles. It then describes in detail the construction of the language and social relation-based factor graph model to address the problem. In addition, a method is also introduced that tries to import more location information about the SNS users by exploiting external Web resources and aims to further improve user language inference. Finally, this chapter describes the construction of a dataset by using the LinkedIn profiles and analyses experimental results.

Chapter 4 formally defines the problem of inferring a user’s topical expertise by exploiting the user’s posted social content, with a focus on Twitter. It then describes in detail the sentiment-weighted and topic relation-regularized learning model to address this problem. After that, it describes the construction of a benchmark dataset of Twitter users versus expertise topics by using Quora profiles. Finally, this chapter analyses and discusses the related experimental results.

Chapter 5 describes in detail the construction of a learning model that infers a user’s topical expertise by exploiting the user’s multiple types of data on a SNS. Again, focusing on Twitter, it then presents the performance of user expertise inference approaches relying on only a single type of user data on Twitter (e.g. tweets, friends, followers or lists) on the dataset from Chapter 4. Finally, this chapter analyses and discusses the experimental results of the proposed model versus other baseline approaches that exploit multiple types of user data on the same dataset.

Chapter 6 presents an approach that matches potential answerers to new questions on a CQA site based on the topical expertise inferred from their social content. It also describes the construction of a dataset by exploiting answering data of Quora users, which is used to simulate the answerer finding scenario on a CQA site. Then, the proposed approach and a conventional approach for answerer finding (i.e. user answering history based) are experimented with on the dataset. The performance of the two approaches are compared and analysed at the end of this chapter.

Chapter 7 concludes the thesis, and discusses the contributions of the PhD research as well as potential future work.
2. State of the Art

This chapter presents the state of the art in user expertise modelling and introduces some other techniques applied in this PhD research. It first reviews existing studies on user expertise modelling, specifically in a SNS setting. However, most studies in the literature are targeted at modelling the expertise of employees in enterprise settings, or expertise of users in question answering communities. They are related to this research in terms of the commonly used modelling methodologies, i.e. the mathematical models used to model a user's expertise. Therefore, this section then reviews the general expertise modelling methods regardless of the scenario in which the user's expertise is modelled. Finally, related work in the area of answerer finding on CQA sites and some other techniques applied in this research are reviewed.

2.1. User Expertise Modelling in Social Networking Sites

Owing to the public availability of user data on Twitter, most existing studies in the field of user expertise modelling in SNSs adopt Twitter as their target research platform. Similarly, experiments conducted in this PhD research are primarily based on the Twitter platform (specifically in Chapters 4, 5 and 6). While the proposed models are general approaches which can be easily adapted to other SNSs, the validation of this is outside the scope of this thesis. Therefore, this section first places emphasis on reviewing the related work of user expertise modelling on Twitter and then other related research targeting other SNSs.

2.1.1. User Expertise Modelling on Twitter

There are three main groups of studies on Twitter that are related to this research: user influence modelling, user expertise modelling on latent topics, and user expertise modelling on specific topics.

2.1.1.1. User Influence Modelling

From the start of the SNS analysis, understanding the influence of users has always been at the core of research in this field. This is because this knowledge about users can benefit a variety of online applications, such as online marketing [Ke15] and information propagation [Go06].
In an early study [Kw10], the authors directly applied the PageRank algorithm [Pa99] to the “following” network on Twitter, where a user corresponds to a node and the following relationships between users correspond to directed edges. The top ranked users are considered as the most influential users. It was found that influential users identified by PageRank are similar to that identified by the number of followers of the user. M. Cha et al. [Ch10] compared the use of three Twitter metrics: indegree, retweets and mentions in measuring user influence on Twitter. They discovered that the number of retweets and mentions serve as better metrics than the number of followers in user influence measurement. There are a number of works [Ha11, Hu13, Ya10, Ja12, Pu12] that extended the PageRank algorithm and incorporated various metrics defined from Twitter to study user influence on Twitter. For example, in work [Ha11], the authors proposed a metric called Magnitude of Influence (MOI) to estimate an initial influence value of each user, which is measured by the size of feedback received from the user’s followers to his/her posted tweets, such as the number of retweets, mentions and favourites. Based on the PageRank algorithm, the MOI values of all the users are propagated recursively along the following network to form the final influence of each user. Huang et al. [Hu13] posed an assumption that the more diverse the people a user can influence are, the more influence the user would gain on the network. They proposed two schemes to evaluate the social diversity of users, which are based on the network structure and the pattern of influence spread separately. By considering the transition probability of social diversity of users over the following and retweeting network, the PageRank algorithm was extended to identify the diversity-dependent influencers on Twitter.

Apart from the PageRank approach, there are a range of other approaches [Pu12, Ra14, Zh16] that have been used to measure the influence of users on Twitter. The authors in work [Pu12] also applied the popular HITS algorithm [Kl99] on the social network Twitter to identify influential users, where the users are connected by their retweeting, replying and mentioning behaviours. Razis et al. [Ra14] considered the H-index metric [Hi10] of the retweet counts and favourite counts to tweets a user posted for user influence estimation. To address the problem of “sybil attacks”, Zhang et al. [Zh16] applied the weighted eigenvector centrality [Bo07] in the network (constructed through retweeting, replying and mentioning connections) to find the true influencers. It was believed that non-sybil users tend to be more careful and selective in retweeting, replying to and mentioning other users.
The works on user influence analysis on Twitter which are reviewed above, focus on modelling the general influence of users in the social network regardless of topic. There is also a group of studies that try to model the influence of users on different topics. In these cases, the content of the user’s tweets usually needs to be analysed in order to correlate the user with specific topics. To model the problem, the Topic-sensitive PageRank (TSPR) [Ha02] algorithm is one of the most straightforward solutions [We10, Ca14, Li14a]. Weng et al. [We10] identified the existence of homophily in the user following relationship on Twitter, i.e. users tend to “follow back” followers who share an interest in common topics. So they proposed a TSPR-like algorithm named TwitterRank to model the influence of users on latent topics. TwitterRank first employed a Latent Dirichlet Allocation (LDA) model [Bl03] to identify the topics the users are interested in, which are then used to construct a topic-specific relationship network between users. Finally, the influence of a user on a topic is measured by a variant of the PageRank algorithm which considers both the topical similarity between users and link structure over the following network. Cano et al. [Ca14] applied existing entity extraction and topic detection services to correlate a tweet to its corresponding entities and topics, which are then used to generate semantic profiles of Twitter users. Based on the TSPR algorithm, the authors proposed an approach called Topic-Entity PageRank that estimates the user’s topical or entity influence by using the semantic profiles of users on the retweeting network.

Furthermore, there exist other solutions [Li14a, He14, Ka15, Le15, Ra15] that have been proposed to model user influence on specific topics. In work [Li14a], the authors considered the quality of the user’s posted tweets to determine the influence of users on specific topics. Metrics defined from retweeting behaviours were used to measure the quality of a tweet. A disadvantage of this approach is that the topics of the candidate users are predefined. Herzig et al. [He14] considered the two roles of users as content authors and readers on Twitter and gave a new interpretation of the notion of user influence. They believed that the influence of a user is determined by their ability to author content that can consistently satisfy the information needs of readers. An Author-Reader Influence model was proposed to model user influence on a topic, in which the user’s influence was estimated by the probability of the user’s content being read and replicated by others (retweets, mentions and replies). The authors in [Ka15] employed Supervised Random Walks [Ba11a] to find influential users on Twitter, in which the similarity between users’ authored content is used to relate them to different topics. Lee and Lim [Le15] considered
the sentiment change when a user diffuses sentimental content to her/his followers to measure user influence. Supervised learning was applied in work [Ra15] to select characteristic attributes that can distinguish influential users from others on a specific topic. It was concluded that the following relationships are the most useful property for influencer identification rather than the interaction relationships. In this work, topics were related to users by the hashtags contained in their posted tweets.

2.1.1.2. User Expertise Modelling on Latent Topics

There is a group of studies that seeks to learn latent factors that could be descriptive of the user from Twitter. Those works that aim to identify users who can produce relevant and high-quality content are particularly considered as related to this PhD research. This is because the learned latent factors of users can be seen as a kind of representation of user expertise. It is also noted that this latent representation of users may also contain other information about the user, such as user preference or interests, based on the different application scenarios. But in reviewing related works, emphasis will be placed on how the user expertise information is modelled from their historical behaviours.

Applying the topic modelling approach to the social content of users, e.g. tweets, is a commonly used way to learn a user’s latent factors [Ra10, Wa12, Ho10b]. For example, authors in work [Wa12] run LDA over several different types of user-related data on Twitter, e.g. tweets, retweets and lists, to learn the latent factors of users, which are represented as probability distributions over latent topics and are taken as a representation of the user’s expertise. They are then used to compare with the user’s actual expertise topics, aiming to explore the usefulness of each type of user data on Twitter for inferring the expertise of users. Since these latent representations of users are not directly understandable to a human, they are mostly utilised to serve the application of personalised recommendation.

Due to the problem of information overload on Twitter, it is challenging for users to find their desired information. One of the most effective solutions to address this problem is to recommend relevant and high-quality information sources, i.e. followees (people to follow), to the information seeker [Ar12]. In order to do so, Matrix Factorization (MF) [Ko09] has often been used to learn the latent factors for target users and followees by factorizing the observed user-followee rating matrix. Note that in reality there are usually no specific user-followee rating scores, and these are often obtained from the implicit feedback of users’ interactions, such as the existing following relationships. The latent
factors of followees can be seen as a representation of a user’s ability to produce a certain level or quality of content on particular topics. In comparison, the latent factors of a target recommendation user (i.e. the user to whom the followees are recommended) represents the user’s level of preference for particular topics. Therefore, these learned latent factors can be used to predict who would be most likely to produce relevant and high-quality content for the target users, i.e. identify potential followees. To better learn the latent factors of potential followees, previous works [Ch12a, Yu14, Gu17, Ch12b, Ma14] consider various information about the target users, the followees and their interactions in the process of factorization. Chen et al. [Ch12a] incorporated the category information of users, such as age, gender or taxonomy, in MF and argued that users who belong to the same class tend to have close latent factors, i.e. similar topics. In [Yu14], the structure of the social network was exploited to constrain the MF model through regularization. Structural characteristics of social networks, such as transitivity (i.e. users tend to follow followees of their followees) and similar sources (followees of the same user tend to have similar features), have been shown to be effective in improving followee recommendation. Recently, Feltoni Gurini et al. [Gu17] proposed to take the temporal alteration of the user’s attitude (represented by three dimensions: sentiment, volume and objectivity) into consideration in the MF model. It was assumed that users who discussed a topic during the same time frame were more similar than users who discussed the same topic during different time frames.

In addition to followee recommendation, there are several works [Lu15, Ge16] that considered the application scenario of location-aware expert recommendation on Twitter, in which latent factors of expert candidates are learned by incorporating the user’s location information. In [Lu15], the authors stated that users may have geo-spatial preferences when seeking expertise. It motivated them to define several location-related factors that could influence the user’s expert selection, such as, (i) the user’s preference for experts on a specific topic may vary based on the location of the user (named region-based locality) and (ii) the spatial preference of users may vary based on different topics (named topic-based locality). These defined factors are characterised through regularization in a MF model and help to learn better latent factors of both users and experts, which are then used to personalise the user’s recommendation experience. Ge et al. [Ge16] extended the two-dimensional representation (user-expert) of user preference to a three-dimensional representation (user-expert-topic), which is modelled by a taper factorization approach. Apart from the geo-spatial information, they hoped to take into
consideration more contextual factors, e.g. topical and social preferences of users, for personalised expert recommendation. It should be noted that both works introduced here considered the scenario of personalised expert identification from expert candidates on a specific topic, where the specific expertise topics of expert candidates are provided beforehand.

### 2.1.1.3. User Expertise Modelling on Specific Topics

The focus of this PhD research is to learn the user’s expertise on specific topics by exploiting their various social behaviors. In the literature, there are a number of works that focus on a particular area/topic and try to find experts in that area or topic.

Bar-Haim et al. [Ba11b] attempted to locate stock experts by exploiting stock-related tweets and assist investors in making trading decisions. In their proposed approach, stock experts were identified based on the consistency between the user’s prediction of stock prices in her/his tweets and the actual change of the stock prices. Specifically, they first obtained the bullish tweets\(^5\) of a user by using a classifier which predicted the rise of certain stocks. Then the prediction of each bullish tweet from the user was compared with the price change of the stock in the next day.

Abbas et al. [Ab15] proposed a method to identify health experts for a given disease from Twitter. Firstly, the keywords related to a disease were extracted from the WordNet database [Mi95] and the HITS algorithm was employed to select the initial expert candidates based on the usage of these keywords in the user’s tweets. Then, a metric defined from multiple criteria, such as the total health related tweets and sentiment of the users in replies to the tweets of expert candidates, was used to find the final experts for the given disease. Also, a cloud infrastructure was utilised to host the huge tweet repositories and deliver the proposed service, which has been shown to greatly improve its scalability. It was noted that the proposed approach focuses more on the modelling of the influence of identified users, instead of deep relation mining between Twitter users and a given disease. This is because the proposed approach was based on the HITS algorithm and the relationship between a disease and an expert mainly relies on word matching schemes.

---

\(^5\) The tweets have a positive outlook on certain stocks
In [Pa11], the authors adopted a manual method to predefined several target expertise topics (three topics: iPhone, oil spill and world cup, are used in this work), and the candidate Twitter users associated with a topic were selected through substring matching between the user’s tweets and the topic name. They also defined a number of features that may account for a user's topical information by using the user’s posted tweets and her/his one-hop following network (i.e. the user’s direct followers and followees), such as the similarity of two successive tweets of the user and the number of hashtags used in the user’s tweets. A probabilistic clustering approach was at last proposed over the feature space to generate a final list of the top authoritative users for a given topic. Because there was no ground truth harvestable, the authors conducted a user study to evaluate the proposed model.

In [Ho16], the authors assumed that experts on Twitter will behave differently from non-expert users. They intended to identify the discriminative behaviours of expert users and hoped to use them for expertise location on Twitter. To achieve this goal, three groups of features from the tweets or retweets posted by users were defined: behavioral features (e.g. the number of followers and the number of hashtags per tweet), linguistic features (e.g. the user’s lexical coherence and frequency of least common word used in the tweets) and stylistic features (e.g. the number of question marks used and the number of sentiment terms used). Then they harvested expert users on four specific topics, Science, Technology, Health & Fitness and Business, by manually selecting related webpages in which expert Twitter users are recommended on certain topics. Finally, logistic regression was employed to distinguish expert users on the four topics by using the three different groups of features. Through analyzing the experimental results, a number of important findings on expertise discovery in the Twitter network were summarized at the end of the work, such as, experts tend to be older Twitter users than other users, experts use significantly more specialized language, and experts tend to retweet posts with more complex sentences. However, the authors did not convert those findings into effective expertise identification approaches and only gave the conclusions based on some statistical results. Their usefulness in the construction of advanced approaches and application in other topics remains unclear.

There are a number of works that attempted to extend user expertise modelling on Twitter to a larger spectrum of expertise topics.
Wagner et al. [Wa12] was the first to propose to utilise the information from a third-party platform (i.e. Wefollow\textsuperscript{6}) to obtain expertise information of Twitter users. In Wefollow, a Twitter user can register a maximum of 3 topics as her/his familiar areas, and for each topic, the registered users are ranked based on the number of followers in the same area. This offers an effective way of building the correlation between a Twitter user's social activities and her/his expertise topics. This study conducted a series of experiments to explore the ability of different types of user data (i.e. tweets, retweets, biography and list data) to predict the user's expertise in certain topics, which include a user study, topic modelling [Bl03] and classification algorithms [Fo05] to learn expertise topics. The experimental results consistently showed that the approaches based on the user's list data perform the best. Although this work contributes to demonstrating the variance of different user data in their ability to predict a user's expertise, they did not explore effective approaches to improve the prediction accuracy based on their findings. It was noticed that the employed classification algorithms obtained a prediction accuracy of below 50\% on some topics. Besides, it can be easily spotted that most of the top ranked users on Wefollow are either public figures or accounts of public organizations and companies. It is significant to investigate effective methods of inferring expertise information of general Twitter users.

In [Gh12], the authors proposed exploiting the list information of Twitter users to infer their topical expertise. They observed that Twitter users create lists to include other users who are considered as experts in a certain topic, and the list metadata (i.e. list name and list description) provide valuable semantic information regarding the topical expertise of the users included in the list. Thus, this work first used a number of text processing techniques to analyse list metadata, and then took unigrams and bigrams of the refined texts as the topics of expertise of the users in the list. Comparative experiments showed that the proposed approach achieved relatively high performance utilising the list metadata alone. Obviously, this approach has a significant drawback, i.e. its inference relies upon the creation of sufficient lists which include the target user, and on the judgement of the list creators regarding the expertise of the users in the lists. However, a large proportion of Twitter users are not included in many, or any, lists in the first place, which results in a direct failure of the proposed approach.

\textsuperscript{6} http://wefollow.com
In [We16], the problem of expert finding on specific topics in Twitter was taken as a search problem, in which each term in the query is treated as an expertise topic. The authors proposed to look at two aspects of users to estimate their level of expertise on a given topic: global authority and local relevance. To model the global authority of a user on a topic, the authors proposed a semi-supervised graph-based probabilistic method (named SSGR) that exploits three different relations (i.e. user-follower, user-list and list-list relations) on Twitter to obtain the global expertise scores of users on the topic. To be more specific, the three relations were taken into consideration in SSGR by employing a normalized Laplacian regularization term, which is subject to the supervised information derived from Twitter crowds, i.e. the loss term to ensure the global authority of users. It should be noted that the global authority scores of users were calculated offline. The local relevance of users on a topic was estimated by a Gaussian-based method, where a user’s posted tweets, bio and the meta-data of lists which include the user are concatenated to form a document of the user. Finally, given a topic query, the candidates were ranked based on the global authority scores and local relevance scores of query terms. In the experiments, 28 sample queries, e.g. “travelling” and “Boston Marathon bombings” were selected to evaluate the proposed approach, where ground truth experts of these queries were obtained through manual labelling. There are two main deficiencies of the proposed approach. Firstly, the assumption that taking every query term as an expertise topic is not reasonably sound. Secondly, the proposed IR-based approach heavily relies on term-matching to find expert candidates for a given query topic. However, it has been shown from our experimental analysis (see Chapters 4 and 5) that frequent use of topic terms does not necessarily indicate that the user has a good level of knowledge about that topic.

In [Ch14], the authors also modelled the problem of expert finding as a search task, but with the aim of finding expert candidates who not only have expertise in a topic but also are well recognized by the local community (called local experts). An expertise framework named LocalRank was proposed to estimate a Twitter user’s local expertise in a query topic based on two aspects of information about the candidate: the user’s topical authority and local authority. The LocalRank leveraged Twitter list data as the inference source and estimated the user’s local authority by exploiting the geo-spatial information embedded in Twitter lists. The intuition behind the proposed method is that the local authority can be reflected by the distance between the core audience of the expert candidate and the query location. To model the topical authority of users, a language model based approach was proposed that propagates the expertise of users along three
different types of connections with the candidate, i.e. user friendships, list-labelling relationships and list-peer relationships. It should be noted that expertise of each Twitter user is initially represented as a vector of terms, in which the terms are extracted from the meta-data of lists that include the user. In the experiments, 56 topic-location queries, such as “technology” in Chicago and “barbecue” in New York, were selected to evaluate the proposed framework, whose ground truth experts were derived from a crowdsourcing strategy.

Li et al. [Li14b] investigated the problem of expertise modelling in terms of a user’s knowledge on a place or a class of places, usually called points of interest (POIs), e.g. a restaurant or a book shop, instead of general expertise topics. They wanted to locate users who have good knowledge of POIs by exploiting the user’s location tracks revealed from their geo-tagged tweets. Three properties of users were used to estimate the user’s expertise on a POI: within-topic activity (interactions the user had with the target POI), within-topic diversity (interactions the user had with other POIs in the category of the target POI), and recency (recent interactions the user had with the target POI). A user survey conducted by the authors also indicated that the selected three properties are considered as the top criteria when assessing a user’s expertise on a POI. Due to the difficulty in collecting ground truth results, the approach was only evaluated with limited testing samples through user trial and no learning based approach was built based on the corresponding findings.

As mentioned in the Section 2.1.1.2, there are several works [Lu15, Ge16] that tried to recommend experts in specific topics on Twitter to online users. However, they focus on personalised recommendation and assume that the expertise topics of users are provided in advance. For each topic, the importance of experts in that topic is re-weighted based on the preference of an individual. Conversely, the focus of this PhD research is to build the initial relations between expertise topics and Twitter users by exploiting their various social activities.

2.1.2. User Expertise Modelling in Other SNSs

As most other popular SNSs, such as Facebook and LinkedIn, do not make their user data publicly available, only limited studies have attempted to mine expertise information in these SNSs. Below, we highlight those that are most closely related to this PhD research.
Duchateau et al. [Du11] underlined the problem of finding the right person to whom one should communicate with given a question, and they observed that a user tends to contact acquaintances in their social networks rather than strangers when they have a question. For each user, they integrated her/his historical activities on multiple social networks (Delicious, Flickr and StumbleUpon) to construct an enhanced user profile of her/him. Given a query (question) issued by a user, a similarity score is obtained between the query and the profile of her/his friends, ranking them in descending order of their ability to answer the query. In this work, terms in the profile were structured by two different strategies: cluster-based and tree-based, which leads to two different approaches to similarity computation. However, in the experiment, the proposed approaches were only tested on a small dataset (320 users), and there was no effective evaluation strategy proposed. Bozzon et al. [Bo13] tried to solve a similar question but targeted the entire population of social networking sites given an expertise need. In the work, a SNS user is associated with a set of resources which include textual information directly related to the user (e.g. tweets a Twitter user posted), or indirectly related to the user (e.g. pages which friends of a Facebook user liked). Then every textual resource is represented as a combination of a set of words and a set of entities after a series of text pre-processing steps, such as named entity recognition and disambiguation. Finally, a weighted strategy was designed to compute a relevance score between an expertise need and every resource based on their word and entity vectors, which results in a ranked list of candidate experts. The authors conducted a user trial on 40 active users of three SNSs (i.e. Facebook, Twitter, LinkedIn) and the experimental results suggested three main findings: 1) Resources indirectly related to the user can help to enhance the ranking precision; 2) Twitter appears to be the most effective SNS for expert finding; 3) Different SNSs appear to perform differently in different domains. However, this work also suffers from two problems: 1) It requires a manual process of parameter configuration in order to achieve satisfactory performance; 2) The proposed approach mostly relies on direct word/entity matching in expert finding but ignores the semantic meaning of the expertise need and resources, and the serious noise in social content.

Popescu et al. [Po13] attempted to mine the potential domain expertise of Pinterest\(^7\) users. They first defined a set of features that can reflect a user's expertise based on their historical pinning activities and those features were then used to identify potential experts.

\(^7\) www.pinterest.com
for four popular Pinterest topics through two common models: a generalized additive model and a generic model. Experiments conducted on a manually labeled dataset showed that the prediction accuracy of these models could reach as high as 70% on some topics. The critical challenge that faced the work was the shortage of labeled data. It is almost impossible to manually label training samples for tens of thousands of expertise topics.

2.1.3. Section Summary

The research on user influence modelling, described in this section, is related to this PhD research as both areas of research are trying to model a relationship between the user and different topics or areas from the users’ social activities (referring to the topic-sensitive influence modelling). However, the former places emphasis on modelling to what extent a user’s actions could influence/affect the actions of other users in the network. Thus, these existing works majorly rely on graph-based approaches. In comparison, this PhD research focuses on modelling to what extent a user’s expertise is reflected by their various social activities. This connection between a user and her/his expertise is more related to activities directly conducted by the user themselves, rather than the activities of other users. Thus, in contrast to studies on user influence modelling, a deeper understanding of the content of the target user is expected for user expertise modelling.

Although latent topics offer an alternative way to represent the expertise of users in SNSs, this PhD research concentrates on the modelling of user expertise in human-interpretable specific topics. In contrast to previous works that rely on rich activities of users for expertise modelling, this PhD research first looks at the use of limited information about cold-start users to attempt to model their expertise. Specifically, this PhD research proposes to exploit the static profiles of users to learn what languages they speak (Chapter 3).

Furthermore, instead of targeting a particular topic/area or a category of expertise topics, this PhD research also attempts to tackle the problem of modelling the user’s general expertise topics from their various social activities, regardless of the geographical location of users (Chapters 4 and 5). Existing research views this problem as a search task and builds modelling approaches based on the assumption that the frequent use of topic terms indicates the user has a good level of knowledge about that topic. Experiments from this PhD research show that this assumption introduces significant noise into user expertise modelling and results in low prediction accuracy. Additionally, existing
research mainly relies on manual labelling to obtain training/testing samples of a certain expertise topic, which restricts research in this area to small sets of expertise topics. To overcome these challenges, this PhD research proposes an approach to automatically discover expertise topics and annotate the expertise topics of SNS users by exploiting available Web resources (described in Chapter 4).

In addition to the underlying textual information related to the user, this research explores several important hidden factors in SNSs that could help user expertise inference, such as the sentiment intensity of the user’s posted content, the semantic relatedness between expertise topics and the connections between different types of data associated with the user. Based on supervised learning, this research proposes novel learning models that incorporate these factors in the process of user expertise inference, with the aim of improving the inference accuracy.

2.2. Expertise Modelling Methods

This section reviews general expertise modelling methods that have been proposed by researchers where expert finding in an enterprise or CQA site setting is the area of focus. For clarity of presentation, the proposed models are classified into two main categories: Generative Probabilistic Models (GPM) and Discriminative Probabilistic Models (DPM) based on the fundamental methods applied. Before getting into a detailed introduction of the models, the notations that will be used in the below subsections are explained.

\[ t \in T, \ w: \ t \text{ is an expertise topic which could be predefined by the system or specified by an expertise seeker at the time of having an expertise need. A topic } t \text{ could be described by multiple terms } w = w_1, w_2, ..., w_n. \text{ } T \text{ is the set of all expertise topics.} \]

\[ u \in U, \ d \in D: \text{ } u \text{ is a user in a system who has expertise in a number of topics, and } U \text{ is the set of all users. } d \text{ is a document that is associated with a user } u. \text{ } D \text{ is the set of all documents in the system.} \]

2.2.1. Generative Models

Due to its positive empirical performance, and flexibility in incorporating various extensions, the generative probabilistic model is one of the most commonly used approaches for discovering the expertise of online users [Po98]. The premise is to estimate the expertise level of a user \( u \) with respect to a topic \( t \) by looking at the probability
\( P(u|t) \), the likelihood of user \( u \) having expertise on topic \( t \). There are two primary ways of estimating the probability, which leads to two categories of models: Candidate Generation Models (CGMs) and Topic Generation Models (TGMs) [Fa07].

### 2.2.1.1. Candidate Generation Models

Candidate Generation Models assume that an expert candidate is generated by a probabilistic model based on an expertise topic \( t \). Thus it directly computes the probability of obtaining the expert candidate \( u \) from the model through the topic \( t \), \( P(u|t) \), which is usually factored as follows [Ca05]:

\[
P(u|t) = \sum_d P(u, d|t) = \sum_d P(u|d, t) \ P(d|t) \tag{2.1}
\]

where \( P(u|d, t) \) denotes the extent to which a user is associated with a document given an expertise topic. In the estimation of this probability, it is usually assumed that expertise topic \( t \) and user \( u \) are conditionally independent given \( d \), irrespective of the information from the topic \( t \). It means the condition probability \( P(u|d, t) \) can be simplified to \( P(u|d) \), i.e. the extent to which a user is associated with a document [Fa07]. \( P(d|t) \) denotes the probability that document \( d \) is relevant to topic \( t \). This probability is usually estimated using the general language modelling method. By using Bayes' Theorem, it can be refactored as follows.

\[
P(d|t) = \frac{P(t|d)P(d)}{P(t)} \propto P(t|d)P(d) \tag{2.2}
\]

where \( P(t) \) is the probability of a topic. As an expertise topic is given, \( P(t) \) is a constant which can be ignored. Because a topic could consist of multiple terms, the above equation can be further factored as follows:

\[
P(t|d)P(d) = \prod_{w \in t} P(w|d)^{n(w,t)} P(d) \tag{2.3}
\]

where \( n(w,t) \) is the number of times that the term \( w \) appeared in \( t \); \( P(d) \) is regarded as a prior on document \( d \) in practice which can be exploited to favor some types of documents, e.g. email. But mostly, \( P(d) \) is considered as uniform, which leads to:

\[
P(u|t) = \sum_d P(u|d) \prod_{w \in t} P(w|d)^{n(w,t)} \tag{2.4}
\]
Thus, the key to estimating the probability $P(u|t)$ is obtaining the association information between a user and a document, i.e. the estimation of $P(u|d)$, which is also the key component in the other expertise discovery models where user-document associations are required. Below, we give some detailed discussion on how previous studies estimate this probability in different application scenarios. In general, it includes two steps: 1) Identify the existence of the association between users and documents; 2) Estimate the strength of these identified associations.

In different application scenarios, users may have different ways of creating and interacting with documents, which leads to various strategies of identifying which users are associated with which documents. There are common cases where this association information is explicitly specified in the document, e.g. authors of academic publications [Fa08], senders and recipients of emails [Ba06b], members of projects [Pe08], posted tweets of Twitter users [Pa11]. Thus, the user-document associations can be directly extracted which suggests a user has knowledge on the expertise topics contained in documents associated with them. There are also cases where this explicit association information is not available. For example, the CSIRO Enterprise Research Collection (CERC)\(^8\) at the Enterprise Track of the TREC 2007 and 2008 provides no explicit user information but only a unified format of email addresses [Ba07]. In these cases, approaches are needed to recognize users from documents based on their identifiers (e.g. email address, person name), which is usually considered as a named entity recognition problem. A simple but practical approach that has been commonly used in previous studies is to model this as a general information retrieval problem, where the user's identifier is modified and taken as the query terms, and all the documents are built as the search corpus. The retrieved documents are then considered to be associated with the user corresponding to the identifier [Pe07, Pe08]. Another common way of discovering the hidden user-document associations is rule-based approaches [Ba06a]. The user identifier is modified and generated to various variants (e.g. possible abbreviations of a person name [Zh06]), which are then used to match against documents using different algorithms e.g. the Aho-Corasick algorithm [Ah75].

The second step is to estimate the strength of the identified user-document associations. The simplest approach for doing this is taking them as binary associations, but this obviously will result in the loss of evidence for expertise discovery (e.g. occurrence

\(^8\) http://es.csiro.au/cerc/
frequency of a user in a document may indicate the expertise strength of this user on the topics in that document [Ba08a]. A more commonly used approach is representing the strength of associations using real numerical scores. For example, in [Ba06a], the association score was computed as the number of the variants of the user identifier which occurred in the documents; Authors in [Fa10] used a supervised machine learning approach to obtain this score, in which, given the training data of expertise topics and relevance feedback of the expert candidates to those topics, the function of the association score is learned by maximizing the associations between expert candidates and documents with respect to the corresponding topics.

2.2.1.2. Topic Generation Models

Topic Generation Models assume that an expertise topic $t$ is generated by a probabilistic model based on an expert candidate $u$. In these models, the $P(u|t)$ is directly factored using Bayes' rule as equation 2.2:

$$P(u|t) = \frac{P(t|u)P(u)}{P(t)} \propto P(t|u)P(u)$$

(2.5)

where, similar to CGMs, $P(t)$ is considered as a constant. Thus, the problem of expert discovery is reduced to the problem of estimating the probabilities $P(t|u)$ and $P(u)$.

$P(u)$ is the prior probability that $u$ is an expert, which allows TGMs to take into consideration the candidate importance in the process of expert finding. Previous studies have shown that considering this prior can boost the performance of TGMs [Fa07, Ba08b]. For example, in a research institute, senior professors are more likely to have relevant expertise on topics covered by the research institute. In this case, if there is not enough evidence available to help find an expert on a given topic, it is reasonable to infer that the important users (e.g. senior professors) have a higher probability of knowing about the topic. There are two main ways of estimating $P(u)$. The first one is the direct application of the prior knowledge from the dataset. For instance, in [Fa07], a user prior probability is defined as proportional to the occurrence frequency of the user's email address in the dataset. This is based on the intuition that if a user is mentioned frequently by other users, it is very likely that she/he is an important/senior user in this community. The second way of estimating $P(u)$ is through the documents associated with the user $u$ [Se08a]:
\[ P(u) = \sum_d P(u|d)P(d) \]  

(2.6)

where \( P(u|d) \) denotes the user-document associations that are introduced above; \( P(d) \) is a prior on document \( d \).

Then, the remaining task is to estimate \( P(t|u) \), which is the probability of the expertise topic \( t \) being generated by a probabilistic model based on user \( u \). Below, we discuss two major approaches to estimating this probability: Candidate Model and Document Model.

The idea of the candidate model is first, for each user, to estimate a user language model \( \theta_u \) based on the documents associated with the user, which is represented by a multinomial probability distribution over the vocabulary of terms [Ba06a], then to compute the likelihood of topic \( t \), i.e. all the terms in the topic, given the user language model \( \theta_u \):

\[ P(t|u) \approx P(t|\theta_u) = \prod_{w \in t} P(w|\theta_u)^{n(w,t)} \]  

(2.7)

Because a user model \( \theta_u \) may contain zero probabilities, smoothing strategies are usually used to estimate this probability [Ba06a, Fa07].

The idea of the document model is to first model the documents and then estimate the probability of a user having expertise in topics through examining the associations between the user and the documents, which can be represented as:

\[ P(t|u) = \sum_d P(t|d,u)P(d|u) \]  

(2.8)

in which \( P(t|d,u) \) is the measure of a document that supports a user having expertise in a topic and \( P(d|u) \) is the association strength between a user and a document. This model is often used as a baseline approach, since it is easy to implement on the top of an existing IR engine and delivers reasonable performance [Ba12].

### 2.2.2. Discriminative Models

In generative probabilistic approaches that are used for prediction tasks, they model a joint distribution on inputs and outputs and estimate the parameters based on a likelihood criterion. The above introduction indicates that the success of generative models often requires certain strong modelling assumptions, such as the independence assumption between terms, and the independence assumption between a topic and a user given a
document. In contrast, discriminative probabilistic approaches directly model the mapping from inputs to outputs, and estimate the parameters by optimizing the objective loss functions. In practice, discriminative models tend to have fewer assumptions, and in theory it has been proven that discriminative models tend to have a lower asymptotic error when the set of training data increases [Ng02]. Additionally, discriminative models can readily incorporate any features in the modelling process. Therefore, in recent years, discriminative learning has attracted much more attention for research into user expertise modelling.

2.2.2.1. A General Discriminative Learning Framework

Fang et al. [Fa10] proposed a discriminative learning framework to directly model the conditional probability $P(u|t)$ for expert finding. They cast expertise modelling as a binary classification problem, and use a binary variable $r \in \{0, 1\}$ to represent the relevance of a user to a topic i.e. 1 is relevant and 0 is irrelevant. Formally, the framework uses $P_\theta(r|u, t)$ to represent the extent to which user $u$ has expertise in topic $t$ and any form of probability function can be employed to estimate the parameters $\theta$ with the training data. Given the relevance judgement $r_{mk}$ for the training user-topic pair $(u_k, t_m)$, a general conditional likelihood $L$ of the training data can be obtained:

$$L = \prod_{m} \prod_{k} P_\theta(r = 1|u_k, t_m)^{r_{mk}}P_\theta(r = 0|u_k, t_m)^{1-r_{mk}} \quad (2.9)$$

where $M$ is the number of topics and $K$ is the number of users. This model can be learned by maximizing the log-likelihood function, i.e. Eq. (2.9), of the training data.

In [Fa10], two specific methods were proposed to estimate $P_\theta(r|u, t)$, which leads to two different discriminative models. Similar to the document models in section 2.2.1, this first model collects expertise evidence of users from documents. Specifically, given a document $d$, the relevance between user $u$ and expertise topic $t$ depends on two factors: the relevance between $t$ and $d$ and the association between $u$ and $d$. The final relevance score of user $u$ and topic $t$ is obtained by averaging over all documents. Thus, the formal probability function is as follows:

$$P_\theta(r = 1|u, t) = \sum_d P(r_1 = 1|t, d)P(r_2 = 1|u, d)P(d) \quad (2.10)$$
where $P(r_1 = 1|t,d)$ is the probability that a document $d$ matches a topic $t$, i.e. the relevance between the two; $P(r_1 = 1|u,d)$ is the probability that a user $u$ is associated with document $d$; $P(d)$ serves as a prior probability that is generally assumed uniform. Both $P(r_1 = 1|t,d)$ and $P(r_1 = 1|u,d)$ are modelled by a logistic function on a linear combination of features.

Inspired by the finding in machine learning that the geometric mean is better than arithmetic mean in some cases [Ta00], the second model combines the expertise evidence using the geometric mean:

$$P_\theta(r = 1|u, t) = \frac{1}{Z} \prod_d \left( P(r_1 = 1|t,d)P(r_2 = 1|u,d) \right)^{\frac{1}{|D|}}$$  \hspace{1cm} (2.11)

where $Z$ is a normalization factor. The parameters of both of models can be estimated by maximizing the conditional log-likelihood function using Broyden-Fletcher-Goldfarb-Shanno Quasi-Newton optimization [De96], and they share the same computational complexity.

### 2.2.2.2. Learning to Rank

In recent years, the emergence and success of Learning to Rank (L2R) models in the field of information retrieval have attracted much attention from researchers in expertise modelling or expert finding. In document retrieval, L2R models represent documents as feature vectors that reflect the relevance of the document to the query, and automatically learn a ranking model based on the training data. It is obvious that these models can be naturally applied to expertise modelling if we take users as the retrieval target instead of documents and define corresponding features for them. In the literature, L2R algorithms are generally classified into three categories according to the different input and output spaces defined, different hypotheses used and different loss functions employed: the pointwise approach; the pairwise approach; and the listwise approach [Li09].

In [Ya09], the authors employed a pairwise-based L2R algorithm Ranking Support Vector Machine (SVM) [He99] to rank the expert candidates for a given topic. Features, like language model features and various metrics of academic performance, were defined to model a ranking function based on the training data, which can predict the relative order of candidates. Macdonald et al. [Ma11] proposed a learning to rank approach for tasks like expert finding and blog distillation, in which the instances of three factors (document ranking models, ranking cutoffs and voting techniques) were used to generate different features. Then a listwise-based L2R algorithm AdaRank was applied to learn a
ranking model, and experimental results showed that the AdaRank outperforms all the
generative probabilistic methods proposed in the literature. Authors in [Mo11, Mo15]
conducted a series of experiments to evaluate the performance of various L2R algorithms
on expertise modelling which covers examples from all the three classes of L2R
approaches, such as AdaRank, RankNet, Additive Groves, SVMmap. Experiments
conducted on the DBLP dataset showed that the listwise-based SVMmap algorithm
outperforms all other L2R algorithms and the state of the art discriminative models.

2.2.2.3. Other models

Since user expertise modelling can be considered as a ranking task or a classification task,
classic ranking and classification methods have been applied to expertise modelling, such
as voting models [Ma06b], topic models [Ro04], graph-based models [Se08b].

Macdonald et al. [Ma06b] considered expertise modelling as a voting problem. The
profile of a user is a set of documents associated with her/him that indicates her/his
expertise. They then used a ranking of documents with respect to an expertise topic to
rank the candidate users. Specifically, the expertise topic is first taken as a query to
retrieve relevant documents, and then every retrieved document is seen as an implicit vote
for the candidate user whose profile contains the document. Yang et al. [Ya14b] exploited
the voting and discussion information on StackOverflow⁹ to model the user’s expertise
on different topics. They proposed a new expertise metric, named Mean Expertise
Contribution (MEC), that evaluates a user’s expertise on a topic based on three expertise
factors: answering quality, question debatableness and user activeness. A number of data
fusion techniques, such as Reciprocal Rank [Zh03] and CombSUM [Fo94] are also
adapted to aggregate the votes for each user, which leads to a final ranking of users with
respect to the target expertise topic.

As topic models provide an effective way to uncover the hidden thematic structure in
document collections, they have been a hot research topic in machine learning and natural
language processing during the last decade. For example, authors in [Zh08] proposed the
use of probabilistic latent semantic analysis for expertise modelling. They pointed out
that generative language models lack the ability to identify semantic knowledge, which
may result in the failure to find correct users due to the absence of topic terms in the
supporting documents. In this work, a mixture model was proposed that utilises a hidden

---

⁹ www.stackoverflow.com
theme layer to model the semantic relations between the expertise topic and the
documents associated with the user. Thus, the users whose supporting documents share
themes with the given topic will be ranked higher even where there is no corresponding
topic term in the documents. Authors in [Sp13] combined the Latent Semantic Analysis
(LAS) model with a collaborative filtering based method for expert user recommendation
in an expert finding website. In this work, expertise of users was not only represented by
the latent topics generated from the user’s authored content using LAS, but also the latent
factors generated from the user-user interaction matrix using the collaborative filter.

People are connected in nature, so a number of studies proposed the use of social network
analysis techniques for expertise modelling [Zh07a, Do03]. These models are built on the
assumption that a user's expertise can be inferred from the expertise of other users she/he
is connected with. PageRank [Pa99] and HITS [Kl99] are the two most commonly used
algorithms for this class of model. To model a user's expertise, in general, they first obtain
an initial expertise score for each user based on the approaches described above, and then
use graph-based algorithms to propagate the scores of all users and re-rank them.

2.2.3. Section Summary

This PhD research focuses on the use of discriminative learning models for user expertise
inference in SNSs, rather than generative models. This is because generative models
largely rely on the co-occurrence of topic terms in user documents in estimating the user’s
expertise. However, social content is quite noisy, and the frequent use of topic terms does
not necessarily indicate the user has a good level of knowledge about that topic, as
indicated by the experiments conducted as part of this research. In addition, in this
research training/testing data is harvested in the form of user-topic pairs where no
ordering information is obtained (i.e. the extent to which the user has expertise in a topic),
therefore, this research considers the expertise modelling problem as a classification
problem, rather than a ranking problem.

Based on discriminative learning, this PhD research aims to directly model the mapping
from the input of various textual features of SNS users to the output of their expertise in
different topics. To better model this mapping, this research explores several important
factors that could help user expertise inference in SNSs, and takes them into consideration
in designing the objective loss functions. Inspired by previous work in the enterprise
setting where the relationships among users, expertise topics and user documents played
an important role in modelling the user’s expertise, this research also hopes to discover similar relationship information in the SNS setting and exploit them to improve user expertise inference. Particularly, this research reveals that the relationships between expertise topics as well as the relationships between the user and different types of social data can help infer the user’s expertise, and have been incorporated in the process of user expertise modelling.

2.3. Answerer Finding in CQA Sites

2.3.1. Problem Background

In a community question answering site, one of the most important goals is to provide question askers with appropriate answers in the shortest possible time. However, the increasing amount of newly posted questions every day poses a significant challenge for answerers to identify the most appropriate questions. This typically results in a large number of questions remaining unanswered for a long time in CQA sites. To address this problem, the key is to locate the potential answerers who are most likely to provide satisfying answers to given new questions. In this research, this procedure is called Answerer Finding (AF) (it may be also called Question Routing or Question Recommendation in other studies).

The problem of answerer finding can be formally defined as follows: given a new question, it is necessary to locate a list of users in a CQA site who are the most suitable and likely to answer the question. Different approaches have been proposed for AF in CQA sites and each approach can be characterised using three general phases [Gu08]:

- Construction of user profiles from the user’s answering history, representing the user’s topics of expertise;
- Construction of question profiles from the question’s text body, representing the topics related to the question;
- Matching the question profile to all the user profiles.

The matching phase will generate a matching score for each user to a given question, based on which potential answerers are ranked accordingly. Below, different AF approaches will be reviewed that address these issues using different methods.
2.3.2. Answerer Finding Approaches

2.3.2.1. Language Model-based Approaches

In an early study, Liu et al. [Li05] first attempted to apply IR techniques to address the AF problem in CQA sites. They assumed that answerers to a given question are those who answered similar questions in the past. The question text was taken as a query and user profiles built from the user’s answering history were viewed as documents. In this work, the authors used several strategies to build a user’s profile/document from his/her past activities, e.g. the strategy using both question and answer texts of all the user’s answered questions, the strategy using only the question text of one of the user’s answered questions. Then, three language models: query likelihood model [Mi99], relevance language model [La01b] and cluster-based language model [Li04], were applied to rank candidate answerers to a given question. Experimental datasets were constructed from a real-world question answering service and for evaluation, the users who actually answered a question were used as the ground truth answerers of that question. Experimental results showed that the user profiling strategy of using all the user’s answered questions (question text only) gave more consistent performance over different datasets and the query likelihood model tended to deliver better performance than the other two language models. This work demonstrated the feasibility of IR-based approaches for answerer finding in CQA sites. However, the authors only considered the direct application of several existing language models and neglected some important factors in the application scenario of answerer finding, e.g. the prior importance of the users, and smoothing strategies in probability estimation.

Following the fundamentals of the language model, Zhou et al. [Zh09] proposed a framework to incorporate more information in a CQA for AF. The framework consists of two steps. In the first step, an initial probability (or matching score) of a user being an appropriate answerer to a given question was estimated using the user’s past answering activities. Correspondingly, three different language models were proposed in this work for this probability estimation: profile-based model, thread-based model and cluster-based model. These models were constructed on the basis of question threads (a thread represents a question post as well as all its reply/answer posts) and took into consideration different factors that could help expertise estimation. For example, in the profile-based model, the matching score was estimated through two components: the probability of each question term being generated by each thread of a user, and the contribution of the
user to each of her/his threads. The thread-based model took each thread as a latent topic and the matching score was estimated by the probability of each latent topic (thread) generating the question terms, and the contribution of a user to the thread of him/her (i.e. questions answered by the user). In both models, Jelinek-Mercer smoothing [Ch96] was applied to overcome the problem of zero probabilities. In the second step of the framework, “replying” relations between users that were extracted from the question threads were exploited to re-rank the potential answerers (from step one) based on the PageRank approach. It was assumed that if a user replied to another user’s questions, it indicates that the former user tends to be more knowledgeable on the topics of the question.

The traditional language models described above often suffer from the vocabulary mismatch problem due to data sparseness of user/question profiles. To attempt to address this issue, Liu et al. [Li13] proposed a hybrid approach to locate answerers for a target question in CQA sites. Apart from the subject relevance between user profiles and question profiles, the hybrid approach also considered user reputation and authority in a category for AF. Different from previous language model-based approaches, the user profiles were built using question categories, and the user’s answering content as well as quality (e.g. votes) for questions in a category were used to represent the user’s knowledge in that category. This knowledge representation was then used to estimate the subject relevance of a user for a given question of a certain category. A user’s reputation in a category was estimated by the ratio of the user’s answers being adopted as the best answer for each individual question within that category. The authority of users in a category was obtained by applying link analysis to the category-based asker-answerer network.

2.3.2.2. Topic Model-based Approaches

Language model-based approaches primarily rely on the co-occurrence of question words, so they are not able to exploit the advanced semantic information in questions and user profiles for answerer finding [Li10a]. Guo et al. [Gu08] first proposed the use of latent topic models to capture this semantic information for improving answerer finding in CQA sites. In this work, by simulating the user behaviour of question asking and answering, the authors proposed a generative model to discover latent topics of questions and users. Thus, matching scores between users and questions were estimated at both the term-level (BM25 algorithm) and topic-level (a topical matching scheme). Similarly, Liu et al.
Li10a applied an LDA model to learn the latent topics of new questions and users, and then a language model (term-level) was linearly combined with it to estimate the matching scores between a new question and users. Additionally, this work also considered user authority and user activity as a prior to re-rank the potential answerers. It was assumed that the authoritative users tend to provide higher-quality answers and active users tend to answer new questions more efficiently. Qu et al. [Qu09] employed Probability Latent Semantic Analysis (PLSA) [Po01b] on the user’s previously asked questions to discover her/his latent topics. A user-word aspect model was proposed to model the joint probability of a user and a word over all the latent topics. The final matching score was measured by the product of all the probabilities of words contained in the new question.

Yang et al. [Ya13] proposed to jointly model the topics and expertise of users in CQA sites. They argued that no one is expert in all topics and a user’s expertise level should be evaluated with respect to the corresponding topics. A generative probabilistic model, named Topic Expertise Model (TEM) was proposed which, in addition to the textual content of the user’s question/answer posts, also exploited the tagging and voting information in the CQA site to better model the topics and expertise of users. It was assumed that users who use a tag to annotate their questions/answers tend to prefer topics related to that tag and users who receive high votes for their questions/answers are considered as experts in the topics related to that question/answer. The authors of this work also claimed that users who provide answers to users who have a high-level of expertise tend to also have a high-level of expertise, thus, they exploited the network structure to enhance the ranking results from the TEM (PageRank-based approach). However, in work [Ya14a], it was argued that the two types of voting: question votes and answer votes, represent two different meanings and should be treated differently in estimating the user’s expertise. The voting score of questions is meant to measure the clarity and usefulness of questions and the voting score of answers can be used to evaluate the usefulness of the answers to the corresponding questions. Based on this intuition, a generative probabilistic model, named User Topical Ability Model (UTAM), was proposed that models the user’s topic-specific descriptive ability by using their textual content (questions/answers and tags) and voting scores of questions, and models the user’s topic-specific expertise by using their textual content and the voting scores of their answers. Like the previous work, the social links between users were also utilised to further improve the estimation of the matching score obtained from the UTAM in this work.
2.3.2.3. **Classification-based Approaches**

Classification-based approaches position the answerer finding task as a classification problem. They represent user and question profiles as features that could be related to their expertise information, based on which they aim to build mappings from users and new questions to appropriate matches. Modelling the AF problem in this form enables us to exploit information in CQA sites that cannot be utilised in language and topic model-based approaches as they primarily rely on text similarities [Ji13].

Dror *et al.* [Dr11] stated that potential answerers can be driven by multiple factors, such as personal expertise, social reward and game experience, thus, they proposed a multi-channel match model that considers multiple dimensions of data in a CQA site to build the user and question profiles. In the construction of question profiles, three families of attributes were defined from the related data: textual, categories and user IDs. For example, in the family of user IDs, seven roles for a question were extracted, which are: *asker*, *best answerer*, *answerers*, *question voters*, *answer voters*, *best answer selectors* and *question tracers*. These attributes/roles correspond to roles a user may play within a question and the user IDs of those users who played a role were assigned as the attribute value for that attribute/role. In the construction of user profiles, the same attributes as that of questions were defined for each of seven different channels, i.e. the seven roles mentioned previously. For example, attributes for all the questions where a user played the role of *answerer* were combined as the attributes of the user for the channel *answerers*. By matching the attributes of a question and a user, more than 500 features were constructed for a question-user pair, which in the end were utilised to learn the classifier and select appropriate questions for each user. In the experiments, it was discovered that the data from the *best answerer* channel contributed most to the classifier and the Gradient Boosted Decision Tree (GBDT) [Fr01b] achieved the best performance when compared with the other classifiers.

Zhou *et al.* [Zh12a] also considered the answerer finding problem as a classification task. Two groups of features: local features and global features, were defined to capture different aspects of users, questions and user-question relations. The local features were extracted and defined from the current question and/or user itself, such as the length of the title of a question (question feature) and KL-divergence between a question’s textual content and the textual content of all the questions a user answered (user-question relation feature). Apart from the content of the current question/answer, data from other
questions/users were also utilised to define their global features, such as the average question title length, and KL-divergence value between a user’s answered questions and all the questions in the community. These global features were considered as smoothing factors that can help to improve the robustness of the classifier. In the experiments, a support vector machine was employed to learn the classifier, and the results revealed that the KL-divergence between a user’s textual content and the textual content of all questions a user answered is the most important feature in classifying questions.

Ji and Wang [Ji13] also identified a number of statistical features (e.g. the percentage of the best answers from a user’s provided answers) and text similarity features (e.g. the probability of generating a question from the user’s answered questions with a language model or the LDA model) that could be descriptive of the user/question. Additionally, an important assumption was made that is, within a question session, answerers are considered as more knowledgeable users than the question asker. It makes it possible to consider ordering information between users in the process of learning. Thus, based on the ordering information and user/question features, RankingSVM [He00], a pairwise learning to rank approach, was employed to learn ranking models (pairwise classification) for answerer finding. Experiments conducted on a StackOverflow dataset showed that RankingSVM significantly outperformed the SVM-based approach that does not consider the ordering information of users.

2.3.2.4. Other Approaches

In addition to the approaches reviewed above, there are some other studies that attempted to tackle the answerer finding problem in different ways and consider alternative factors to improve prediction accuracy. Zhao et al. [Zh15] found that most users in CQA sites do not have a sufficient volume of past answering activities to construct a user profile and have been ignored in AF. Thus, they considered answerer finding on a CQA site as a missing value estimation problem, i.e. predicting unknown values in a rating matrix by matrix completion, in which the content of questions is considered as additional information. To better predict the missing values in the rating matrix, the social relation between users was used to regularize the process of matrix completion. They assumed that if two users have a strong connection in a SNS, they tend to qualify for answering similar questions. It was also observed that over one third of the users in a CQA site (i.e. Quora) have a Twitter account, thus, the “following” relation in Twitter for these CQA users was extracted to build their social connections. Cheng et al. [Ch17] tackled the AF
problem as a multi-objective ranking problem and tried to simultaneously optimize the answering possibility and answer quality of target users for answerer finding. Based on the learning to rank algorithm LambdaMART [Bu10], a multi-objective approach was proposed through redesigning its smooth approximate gradient. To address the cold start problem in CQA sites, Srba et al. [Sr15] proposed to exploit non-Question Answering (non-QA) information about the user for answerer finding. They were inspired by the success of previous work that showed non-QA data, such as social activities in social networking sites, can reflect the online user’s expertise. Therefore, they used not only a user’s past answering activities, but also non-QA data associated with the user to build the user’s profile. A user’s expertise for a new question was estimated by a linear combination of two aspects of the user: the probability of the user’s QA profile generating the question and the probability of the user’s non-QA profile generating the question. The two probabilities were weighted through a dynamic coefficient as it was assumed that the non-QA profile should play a more important role for inactive users in a CQA site. In the experiment, data from a user’s personal description, homepage and Twitter account was used as non-QA data and results showed that the non-QA data can improve the ranking for both inactive and active users.

2.3.3. Section Summary

This section reviewed a variety of approaches that tackle the answerer finding problem in different ways. In Chapter 6, this PhD research also attempts to address this problem. In contrast to the studies outlined above that primarily rely on the user’s answering history for user profile construction, this research focuses on the scenario of cold-start users, i.e. users on a CQA site that have very few or no previous answering activities. These cold-start users do not have sufficient data for profile construction and they have often been ignored in AF. This PhD research proposes to exploit the expertise information that is inferred from the user’s SNSs (by our proposed inference approaches) for AF. This research aims to explore the potential of the inferred expertise information in this particular application scenario and to include more CQA users into the process of answerer finding. In contrast to previous work that also relies on non-QA data for AF, this research models the expertise of users in specific topics instead of latent topics (i.e. probability distribution over word vectors). The representation of user expertise using specific topics is human understandable and therefore easier to apply in a variety of application scenarios.
2.4. Other Related Work

The previous three sections reviewed the existing work that is related to the core of this PhD research. This section will simply review two aspects of work that is also involved in the research: (1) User language information inference, and (2) Tweet sentiment analysis.

2.4.1. User Language Information Inference

Users' language information is an important input to multilingual applications. This PhD research proposes to use the online user’s social profiles to infer their language information (Chapter 3). The previous sections reviewed related work from two different perspectives: user expertise inference in SNSs (section 2.1) and expertise modelling approaches (section 2.2). However, from the perspective of how to automatically acquire online users’ language information, the previous studies primarily rely on language identification (LID) techniques [Du94, Xi09]. In other words, they infer what languages a user comprehends by identifying what languages of texts the user previously read/wrote. Distinct LID techniques may be used on different types of texts to harvest their language information, e.g. Web pages [Ma05], search queries [St10], social texts [Ca13]. However, to the best of our knowledge, this PhD work is the first attempt to acquire the users' language information via their social profiles. Specifically, the inference is based on the information stored in the profile itself and not on the posts or media shared by the user on their profile’s space (e.g. their “wall” in Facebook). This approach is useful in cold-start scenarios where there are no prior user interactions available to base the LID decision upon.

2.4.2. Tweet Sentiment Analysis

Apart from studies on user expertise inference, this PhD research is also related to tweet sentiment analysis techniques. This is because sentiment analysis is employed to evaluate the importance of textual features in one of the proposed learning models for user expertise inference (Chapter 4). Although this research does not intend to propose a new sentiment analysis approach, it is important to understand the state-of-the-art sentiment analysis techniques and select a suitable one for the proposed model.

In the literature, there are two main categories of sentiment analysis approaches: machine learning based approaches and lexicon-based approaches. Machine learning based
approaches mainly rely upon supervised classification algorithms, e.g. SVM [Sa12a], Maximum Entropy [Sp11] and Naive Bayes [Sa12b], to learn a classifier and identify the sentiment orientation and/or intensity of a tweet. However, this category of approaches often requires extensive training data and is highly sensitive to the domain from which the training data is extracted [Li12]. They do not suit the learning model proposed in this PhD work, because it is very difficult to build a training dataset for each expertise topic. It is not surprising that the sentiment words are the most important indicators of sentiment. Thus, the lexicon-based approach assesses the sentiment orientation of a text based on the sentiment words and phrases in the text. In this category of approaches, a comprehensive, high quality lexicon is essential where rich sentiment words and phrases are assigned with their corresponding sentiment polarity or intensity, based on which different sentiment analysis approaches can be designed with respect to specific application scenarios [Tu10, Ch12c, De13, Hu14]. For example, the LIWC lexicon\(^\text{10}\), that was originally constructed to study the various emotional, cognitive, and structural components present in individuals’ verbal and written speech samples [Pe15], has been used to extract the political orientation of Twitter users [Tu10]; SentiWordNet\(^\text{11}\), built based on the WordNet\(^\text{12}\) ontology [Es07], has been used to analyse customers' opinions on specific products on Twitter [Ch12c]. Hutto et al. [Hu14] constructed a comprehensive sentiment lexicon for sentiment analysis in the microblog-like context, which combined several existing sentiment lexicons, e.g. LIWC, ANEW\(^\text{13}\), and also incorporated several types of special lexical features that are often used to express sentiment in microblogging services, e.g. western-style emoticons\(^\text{14}\), sentiment-related acronyms and initialisms. Additionally, an extensive manual annotation process was then conducted to estimate the sentiment intensity of lexical features. Finally, based on the constructed lexicon, a rule-based sentiment analysis approach named VADER\(^\text{15}\) was developed which takes into consideration a number of grammatical and syntactical conventions when humans express or emphasize sentiment intensity. Experimental results reported in the work showed that VADER achieved remarkable prediction accuracy in tweet sentiment analysis and outperformed eleven other highly related sentiment analysis approaches. Therefore, this PhD work adopts VADER as the sentiment analysis approach in the proposed model and

\(^\text{10}\) http://www.liwc.net  
\(^\text{11}\) http://sentiwordnet.isti.cnr.it/  
\(^\text{12}\) https://wordnet.princeton.edu/  
\(^\text{13}\) http://csea.phhp.ufl.edu/media/anewmessage.html  
\(^\text{14}\) http://en.wikipedia.org/wiki/List_of_emoticons#Western  
\(^\text{15}\) https://github.com/cjhutto/vaderSentiment
leaves the study of the influence of different sentiment analysis approaches on the model as a future work.

2.5. Chapter Summary

This chapter has presented the state of the art on the existing user expertise modelling approaches and other related techniques applied in this PhD research. It first reviewed the existing research on user expertise modelling in a SNS setting and identified possible ways of modelling SNS users’ expertise through the exploitation of their various social activities. However, none of the existing approaches looked at the use of limited information about cold-start users to model their expertise. Thus, this PhD research proposes to exploit the static profiles of users to learn what languages they speak (Chapter 3). In addition, the author of the thesis also identified the deficiencies in exiting works in modelling the user’s general expertise topics from their various social activities (summarised in Section 2.1.3). One of the major problems is that most existing works view this problem as a search task and build modelling approaches based on the assumption that the frequent use of topic terms indicates the user has a good level of knowledge about that topic. Experiments from this PhD research show that this assumption introduces significant noise into user expertise modelling and results in low prediction accuracy in the SNS scenario. This PhD research proposes to design novel modelling approaches to address these problems and improve the prediction accuracy (Chapter 4 and Chapter 5).

In this chapter, a state of the art review on general expertise modelling methodologies, mainly from the literature on employee expertise modelling in enterprise settings or user expertise modelling in question answering communities, was then undertaken. This review was performed to provide the theoretical foundations for the design of user expertise modelling approaches in this PhD research.

The state of the art on answerer finding approaches on CQA sites was also reviewed in this chapter. This review was performed to demonstrate that existing studies primarily rely on the user’s answering history to model user expertise and find potential answerers for newly posted questions. However, there is a large proportion of cold start users on CQA sites who have very few or no previous answering activities. These users have often been ignored by the answering history based approaches in answerer finding. Thus, this PhD research proposes to exploit the expertise information that is inferred from the user’s
SNSs (by our proposed inference approaches) to help answerer finding on a CQA site (Chapter 6).

Finally, this chapter reviewed two aspects of work that are also involved in the PhD research: 1) User language information inference, and 2) Tweet sentiment analysis. As this PhD research proposes to use the online user’s social profiles to infer their language information in Chapter 3, the review was performed to examine the related work on user language information modelling from the perspective of language identification techniques. Also, as sentiment analysis is employed to evaluate the importance of textual features in one of the proposed learning models for user expertise inference in Chapter 4, the review was performed to understand the state-of-the-art sentiment analysis techniques and select a suitable one for the proposed model.
3. Language Information Inference Using Social Profiles

3.1. Introduction

As outlined in Chapter 2, previous research primarily focused on the use of the abundant user-generated content available on SNSs for expertise inference, such as posts authored by a user and information about a user’s friends. However, recent studies [Ba15] have revealed that a large proportion of SNS users are relatively inactive and publish very little content, i.e. cold-start users. This prevents the previously proposed approaches from inferring their expertise information. In this PhD research, it was observed that in many SNSs, new users are asked to provide some basic personal information like education and work experience when registering, such as on Facebook and LinkedIn. These static profiles provide first-hand information about a new user to the SNS and SNS applications. There is a chance that the information in a user's social profile may implicitly indicate their expertise information. In particular, this PhD research proposes to exploit the static profiles of new users to infer their language information. This proposal can be assessed based on: (1) the feasibility of the proposed idea, and (2) the significance of the proposed research.

- **Feasibility**: The intuition is that a user's experiences could imply what languages they know. For example, if a user has conducted academic studies in Germany, then this could imply that the user at least understands German to some extent; if a user lived in multiple places that share a common language, it is reasonable to infer that this user likely comprehends this language.

- **Significance**: As a result of globalization and cultural openness, it has become common for modern-day humans to speak multiple languages (polyglots) [Tu99, Di10]. Knowledge of what languages the online user comprehends¹⁶ is becoming important for many Web applications to enable effective information services. For example, knowing the user’s language information enables: search engines to deliver multilingual search services; machine translation tools to identify optional target translation languages; and advertisers to serve targeted international ads.

Traditionally, there have been two primary ways to acquire an online users' language information: (1) explicitly ask the user to specify this information [Gh14]; (2) leverage

¹⁶ In this thesis, *comprehend* means that the user is able to grasp information in that language to a certain extent.
user centric data (e.g. browsing history) to detect what languages the user comprehends through language identification techniques [Oa09]. However, online users often choose not to explicitly provide this expertise information, even when they have a facilitated means to do so. For example, based on our analysis of 50,575 user profiles on LinkedIn, we found that only 11% of users specified the languages that they speak, even though there are input fields available for this; Ghorab et al. [Gh10] carried out an analysis where it was found that many users of The European Library\textsuperscript{17} entered search queries in non-English languages and browsed documents in those languages, without using the drop-down menu that allows them to specify the interface language. An additional challenge faced by the second approach is the common cold-start problem, where there is only limited history of interactions available for a new user.

This PhD work proposes the use of static social profiles to automatically infer the languages that a user comprehends. It aims to overcome the cold-start problem, and benefit a number of Web applications, as mentioned previously. In particular, Multilingual Information Retrieval (MIR) is a prominent application area for this study, in which information about the user's preferred language and which other languages the user comprehends are considered vital parts of the user model [Gh14]. This work is a first step towards automatically inferring the user's level of expertise in languages that they comprehend. Furthermore, this research can be integrated with other work in user expertise/characteristics profiling, where other aspects of information about the user can be automatically inferred, such as expertise in specific topics (Chapters 4 and 5).

It is straightforward to cast the task of user language inference from social profiles as a standard text classification problem. In other words, predicting what languages a user comprehends relies on features defined from textual information of the social profiles, e.g. unigram features. However, social profiles are usually incomplete, with critical information often missing. For example, the location information of the work experience, which, in this research, is shown to be important evidence for language inference, is only provided by about half of the users in the collected dataset. Moreover, some languages can be mutually intelligible (i.e. speakers of different but related languages can readily understand each other without intentional study) or regionally related (i.e. multiple languages are spoken in one region). These languages may share many common features.

\textsuperscript{17}http://www.theeuropeanlibrary.org
which makes it hard to identify the discriminative features between them. Therefore, solely relying on textual features to infer these languages may yield unsatisfactory results.

To address these challenges, this PhD work investigates three factors to better model the problem: (1) *Textual attributes in social profiles*. These attributes provide fundamental evidence about what languages a user may comprehend. This work also attempts to exploit external resources to enhance the textual attributes. It aims to import more information that is associated with the user and which may also reflect user language information. (2) *Dependency relations between languages*. Languages may be related to each other in certain ways, e.g. mutually intelligible. This relation could reflect the possibility that a user comprehends other languages based upon a language we know the user understands. (3) *Social relations between users through their social profiles*. It is reasonable to assume that users with similar academic or professional backgrounds may comprehend one or many of the same languages. This relation information between users could be extracted from their social profiles. Finally, a Language and Social Relation-based Factor Graph Model (LSR-FGM) is proposed which predicts the user language information under the collective influence of the three factors.

Experiments are conducted on a large-scale LinkedIn profile dataset. Experimental results show that LSR-FGM clearly outperforms several alternative models and is able to obtain an F1-score of over 84%. Experiments also demonstrate that all three factors outlined above contribute to the process of inference.

The rest of the chapter is organized as follows. Section 3.2 explains the challenges and corresponding solutions for inferring language information using social profiles, then formally defines the problem; Section 3.3 details the proposed LSR-FGM and Section 3.4 introduces the collected dataset and gives the experimental results. Finally, Section 3.5 concludes this chapter.

### 3.2. Problem Description

This section first gives some background information on the inference problem and then introduces two main challenges in modelling the problem, as well as the corresponding proposed solutions. Finally, it gives the formal problem definition.
3.2.1. **Background**

In general, a social profile consists of multiple fields, each of which details a particular aspect of information about the user, such as education background, hobbies etc. Different platforms may use different fields to construct user profiles. This work considers three commonly used fields in the profiles of popular SNSs such as LinkedIn, Facebook and Google Plus, for the language information inference problem:

- **Summary**: Unstructured text that gives a general introduction about the user. Because there is no structure restriction, the focus of this field varies from user to user.

- **Education Background**: Structured text that details each study experience of the user by subsections. Each subsection could include attributes like school name, study major, etc.

- **Work Experience**: Structured text that details the user’s work experience by subsections. Each subsection could include attributes like company name, role, work period, etc.

In practice, the language information of some users is already known through certain means. For example, it is explicitly stated in the user's profile; or it is predictable from the user's interactions with the system. Therefore, the problem is how to infer the language information of the remaining users based on the textual information of the given social profiles and the known language information of a subset of the users.

3.2.2. **Challenges and Solutions**

As discussed in Section 3.1, the attributes in a social profile may implicitly suggest what languages a user comprehends, especially for the location-related attributes. For example, a user who lived in a place for a long period of time probably comprehends one of the languages spoken in that place. Based on this assumption, we can model the language inference using social profiles as a supervised classification problem. However, there are two main challenges in modelling this problem:

1. Users' online social profiles are usually incomplete and some profiles even miss critical information as outlined in Section 3.1. Additionally, certain information is generally not asked for by SNS platforms when a user is populating their social profile, but which may provide important evidence for language information inference, e.g. the location of institutions in the Education Background field.
In order to alleviate this problem, this work proposes to correlate each experience of the user to the corresponding location (if missing) by exploiting external resources. For example, a University can be linked to the homepage of this University, from which the location can be harvested. Then, more language related attributes can be imported to help inference. The specific strategy adopted in this work is detailed in Section 3.4.1.

(2) Some attributes in the profile may be misleading in the inference process. As languages could be regionally related or mutually intelligible, they may share similar discriminative features. When many of these languages are taken into consideration in the target language set, only considering the textual attributes as features may not distinguish between these languages.

In this work, Chinese, French, German, Hindi and Spanish are selected as target inference languages. In the five languages, Spanish, French and German are used in combination as official languages in a number of multilingual countries. For example, both French and Spanish are official languages of Equatorial Guinea; French and German are used as official languages of Luxembourg\(^\text{18}\). Also, French has lexical similarities of 0.75 and 0.29 (where the maximum is 1.0: a total overlap between vocabularies) with Spanish and German respectively\(^\text{19}\). By contrast, Hindi and Chinese are spoken as an official language only in India and China respectively; they have no overlap with each other or the above three languages in vocabulary. So lower prediction accuracy is expected on French, German and Spanish (when compared with Chinese and Hindi) as it is more difficult to identify their discriminative features. This intuition is validated in the experiments conducted as part of this research.

It is noted that, apart from Hindi, there are many other languages (which in the context of this research are considered to be related languages of Hindi) which are spoken in India, and that tertiary education within the country is conducted primarily through English. While English and other Indian languages are not considered in the target inference language in this PhD research, the location-related information about the user can increase the probability estimation of the user’s likelihood to have language expertise in Hindi. The introduction of other related Indian languages is expected to decrease the inference accuracy on both Hindi and these languages.

\(^{18}\) http://en.wikipedia.org/wiki/List_of_multilingual_countries_and_regions

\(^{19}\) http://en.wikipedia.org/wiki/Lexical_similarity
This work takes two types of relation into consideration when modelling the language inference problem in an attempt to address this challenge:

- **Dependency relation between languages**: The above example also hints that if a user knows French, she has a much higher probability of also knowing Spanish than Chinese. Thus, this dependency relation between languages could be helpful in inferring users' language information.

- **Social relation between users**: Although new users have no direct friendship/followship connections with other users, they can be related through available information of their social profiles. This PhD work focuses on the *same-experience* relation, i.e. two profiles share a study experience (studied the same major in the same institute) or work experience (worked in the same role in the same company), to help inference. For example, two users who shared the same work experience may imply that they know a common language because communication is needed between employees in the department of the company. Thus, it is reasonable to assume that users with the same-experience relation are likely to know a common language.

Therefore, this work proposes a model that collectively considers the three factors outlined above: enhanced textual attributes, language relation and social relation, to model the problem of language inference using social profiles.

### Table 3-1: Definition of Notations

<table>
<thead>
<tr>
<th>Notation</th>
<th>Notation Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$N$</td>
<td>Number of users/profiles</td>
</tr>
<tr>
<td>$K$</td>
<td>Number of target languages</td>
</tr>
<tr>
<td>$W$</td>
<td>User attribute matrix</td>
</tr>
<tr>
<td>$v = (u_i, l_j)$</td>
<td>A correlation node consists of user $u_i$ and language $l_j$</td>
</tr>
<tr>
<td>$V$</td>
<td>Correlation node matrix</td>
</tr>
<tr>
<td>$X$</td>
<td>Attribute matrix of $V$</td>
</tr>
<tr>
<td>$y \in {+1, -1}$</td>
<td>Has or doesn’t have expertise</td>
</tr>
<tr>
<td>$E$</td>
<td>Social and language relations</td>
</tr>
<tr>
<td>$\alpha, \beta, \gamma$</td>
<td>Model coefficients</td>
</tr>
<tr>
<td>$Z1, Z2$</td>
<td>Normalization factors</td>
</tr>
</tbody>
</table>
3.2.3. Problem Definition

The input of $N$ social profiles can be represented as $G = (U, L, W)$, where $U$ is the set of $|U| = N$ users and $L$ is the set of $|L| = K$ target inference languages; $W$ is an attribute matrix associated with users in $U$ in which each row corresponds to a user, each column an attribute of the profile, and an entry $w_{ij}$ denotes the attribute value of the $j^{th}$ attribute in the profile of user $u_i$. The objective of the work is to learn a model that can effectively infer what languages a user comprehends.

This work defines the correlation node $v = (u_i, l_j)$ that is associated with a label $y$ (binary value) to represent the output of the problem. It means user $u_i \in U$ comprehends language $l_j \in L$ if $y = 1$ or the opposite if $y = -1$. Each user in $U$ is mapped to $K$ correlation nodes with the $K$ languages in $L$, so a set $V$ with $K \cdot N$ correlation nodes and a corresponding label set $Y$ are obtained. As part of users' language information is known, i.e. label values of the corresponding correlation nodes are given, they are denoted as set $Y^L$. The remaining labels are denoted as set $Y^U$, where $Y = Y^L \cup Y^U$. In addition, the correlation nodes are connected through the two types of relations defined above, which constitute an undirected edge set $E$. The definition of $E$ is detailed in the next section. The definition of notations can be found in Table 3-1.

Therefore, given a partially labeled network $G = (V, U, L, Y^d, E, W)$, the objective of the language inference problem is to learn a predictive function: $f: G \rightarrow Y^U$.

3.3. Language and Social Relation-based Factor Graph Model

This section details the construction of our Language and Social Relation-based Factor Graph Model and proposes a method to learn this model.

3.3.1. Model Definition

The LSR-FGM collectively incorporates the three factors outlined above: textual attributes, language dependency relations and same-experience social relations, to better model the problem of user language information inference. Its basic idea is to define these relations among users and languages using different factor functions in a graph. Thus, an objective function can be defined based on the joint probability of all factor functions. Learning the LSR-FGM is to estimate the model parameters, which can be achieved by maximizing the log-likelihood objective function based on the observed information.
Below, we introduce the construction of the objective function in detail. Fig. 3-1 shows the graphic representation of the LSR-FGM.

For simplicity, given a correlation node \( v_i \), we use \( v_i(u) \) and \( v_i(l) \) to represent the user and language of this node respectively. Note each correlation node \( v_i \) in \( V \) is also associated with an attribute vector \( x_i \) which is from the attribute vector of user \( v_i(u) \), and \( X \) is the attribute matrix corresponding to \( V \). Then, we have a graph \( G = (V, E, Y, X) \), in which the value of a label \( y_i \) depends on both the local attribute vector \( x_i \) and the connections related to \( v_i \). Thus, we have the following conditional probability distribution over \( G \):

\[
P(Y|X, E) = \frac{P(X, E|Y)P(Y)}{P(X, E)}
\]

(3.1)

According to the Bayes' rule and assuming \( X \perp E \) in LSR-FGM, we can further have:

\[
P(Y|X, E) \propto P(X|Y)P(Y|E)
\]

(3.2)

in which \( P(X|Y) \) represents the probability of generating attributes \( X \) associated to all correlation nodes given their labels \( Y \), and \( P(Y|E) \) denotes the probability of labels given all connections between correlation nodes. It is reasonable to assume that the generative probability of attributes given the label value of each correlation node is conditionally independent. Thus, we can factorise Eq. (3.2) again:

\[
P(Y|X, E) \propto P(Y|E) \prod_i P(x_i|y_i)
\]

(3.3)
where \( P(x_i|y_i) \) is the probability of generating attribute vector \( x_i \) given label \( y_i \). Now the problem is how to instantiate the probability \( P(Y|E) \) and \( P(x_i|y_i) \). In principle, they can be instantiated in different ways. This work models them in a Markov random field, so the two probabilities can be instantiated based on the Hammersley-Clifford theorem [Ha71]:

\[
P(x_i|y_i) = \frac{1}{Z_1} \exp \left\{ \sum_{j=1}^{d} \alpha_{(v_i(l),j)} \cdot f_{(v_i(l),j)}(x_{ij}, y_i) \right\}
\]

\[
P(Y|E) = P(Y|E_{LANG}, E_{EXP})
\]

\[
= \frac{1}{Z_2} \exp \left\{ \sum_{v_i \in V} \left[ \sum_{v_j \in LANG(v_i)} g(v_i, v_j) \right] \prod_{v_k \in EXP(v_i)} h(v_i, v_k) \right\}
\]

in which, \( Z_1 \) and \( Z_2 \) are normalization factors. In Eq.(3.4), \( d \) is the length of the attribute vector; a feature function \( f(x_{ij}, y_i) \) is defined for each attribute \( j \) (the \( j^{th} \) attribute) of correlation node \( v_i \) for the language \( v_i(l) \), and \( \alpha_{(v_i(l),j)} \) is the weight of attribute \( j \) for language \( v_i(l) \). In Eq.(3.5), \( E_{LANG} \) and \( E_{EXP} \) are edges between nodes in \( V \) through language dependency relations and same-experience relations respectively; two sets of relation factor functions \( g \) and \( h \) are defined which correspond to \( E_{LANG} \) and \( E_{EXP} \) respectively; \( LANG(v_i) \) denotes the set of correlation nodes having the same user as \( v_i \) but with different languages (language dependency relation); \( EXP(v_i) \) denotes the set of nodes in which the users have the same-experience relation with the user of \( v_i \). Next, we will introduce the specific definitions of the feature functions \( f \), and relation factor functions \( g \) and \( h \) adopted in LSR-FGM.

**Local textual feature functions**: The unigram features of the textual information in social profiles are used to build the attribute vector space, and they are also used as binary features in the local feature function for each target language. For instance, if the profile of a user contains the \( j^{th} \) word of the attribute vector space and specifies she knows language \( l \), a feature \( f_{(i,j)}(x_{ij}=1, y_i=1) \) is defined and its value is 1; otherwise 0. This feature definition strategy is commonly used in graphical models like Conditional Random Field [La01a]. Therefore, the conditional probability distribution \( P(X|Y) \) over \( G \) can be obtained:
\[
P(X|Y) = \frac{1}{Z_1} \exp \left\{ \sum_{v_i \in V} \sum_{j=1}^{d} \alpha_{(v_i(l), j)} f_{(v_i(l), j)}(x_{ij}, y_i) \right\}
\]

**Language dependency relation factor:** Any two nodes in \( V \) are connected through the language dependency relation if they are from the same user. If nodes \( v_i \) and \( v_j \) have a language dependency connection, a language dependency relation factor is defined as follows:

\[
g(v_i, v_j) = \beta_{ij} (y_i - y_j)^2
\]

where \( \beta_{ij} \) represents the influence weight of node \( v_j \) on node \( v_i \).

**Same-experience social relation factor:** The nodes in \( V \) are connected through a same-experience relation if users of the nodes share a work or study experience. Similarly, a same-experience relation factor is defined if two nodes have this connection:

\[
h(v_i, v_j) = \gamma_{ij} (y_i - y_j)^2
\]

where \( \gamma_{ij} \) represents the influence weight of node \( v_j \) on node \( v_i \) through the same-experience relation factor.

Finally, LSR-FGM can be constructed based on the above formulation. By combing Eqs. (3.3)-(3.8), we can define the objective likelihood function as:

\[
P(Y|X, E) = \frac{1}{Z} \exp \left\{ \sum_{v_i \in V} \sum_{j=1}^{d} \alpha_{(v_i(l), j)} f_{(v_i(l), j)}(x_{ij}, y_i) \right. \\
+ \left. \sum_{v_i \in V} \sum_{v_j \in \text{LANG}(v_i)} \beta_{ij} (y_i - y_j)^2 \right. \\
+ \left. \sum_{v_i \in V} \sum_{v_k \in \text{EXP}(v_i)} \gamma_{ik} (y_i - y_k)^2 \right\}
\]

where \( Z=Z_1Z_2 \) is a normalization factor and \( \theta = (\{\alpha\}, \{\beta\}, \{\gamma\}) \) indicates a parameter configuration.

Thus, we built the LSR-FGM with the objective likelihood function Eq.(3.9). This model aims to best recover the label values \( Y \), which can be represented by maximizing the objective likelihood function given the observation data.
3.3.2. Model Learning and Prediction

The last issue is to learn the LSR-FGM and to infer unknown label values $Y^U$ in $G$. Learning the LSR-FGM is to estimate a parameter configuration of $\theta$ from a given partially labeled $G$, which maximizes the log-likelihood objective function $L(\theta) = \log P_\theta (Y^L|X,E)$, i.e.

$$\theta^* = \arg \max \log P_\theta (Y^L|X,E)$$  \hspace{1cm} (3.10)

This work uses a gradient decent method to solve the objective function. Taking $y$ as an example to explain how to learn the parameters, first the gradient of each $\gamma_{ik}$ with regard to the objective function $L(\theta)$ can be obtained:

$$\frac{L(\theta)}{\gamma_{ik}} = \mathbb{E}[h(v_i,v_k)] + \mathbb{E}_{P_{\gamma_{ik}}(Y|X,E)}[h(v_i,v_k)]$$  \hspace{1cm} (3.11)

in which, $\mathbb{E}[h(v_i,v_k)]$ is the expectation of factor function $h(v_i,v_k)$, i.e. the average value of the factor function $h(v_i,v_k)$ over all the same-experience connections in the training data; the second term is the expectation of factor function $h(v_i,v_k)$ under the distribution $P_{\gamma_{ik}}(Y|X,E)$ given by the estimated model. Similarly, the gradients of $\alpha$ and $\beta$ can be derived.

As the graphical structure of input $G$ can be arbitrary and may contain cycles, it is intractable to directly calculate the second expectation. A number of approximate algorithms have been proposed to address this challenge, such as Loopy Belief Propagation (LBP) [Mu99] and Mean-field [Xi02]. In this work, LBP is adopted to approximate the gradients considering its ease of implementation and effectiveness [Ta11, Zh12b]. In each iteration of the learning process, LBP is employed twice, once for estimating the marginal distribution of unknown variables $y_i$ and another for the marginal distribution over all connections. Then, the parameters $\theta$ are updated with the obtained gradients and a given learning rate in each iteration.

It is clear that LBP is employed to infer the unknown labels $Y^U$ in the learning process. Therefore, after convergence of the learning algorithm, all nodes in $Y^U$ are labeled which maximizes the marginal probabilities. Correspondingly, the language information of all unlabeled users is inferred.
3.4. Experiment and Results

This section first describes the construction of the dataset and the strategy used to import location information that is associated with the profile from external resources. It then introduces the comparative methods and, finally, presents and discusses the experimental results.

3.4.1. Dataset Construction and Location Information Enhancement

This work constructed a dataset using LinkedIn profiles, as most profiles in LinkedIn are publicly available and a field exists for language information (used as ground truth) and the three fields required (used as inference source) for the inference model are also available. For privacy protection, the names of the profiles are not collected.

Table 3-2: The distribution of languages in the collected social profiles

<table>
<thead>
<tr>
<th>Language</th>
<th>Number</th>
<th>Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>3396</td>
<td>45.5%</td>
</tr>
<tr>
<td>Chinese</td>
<td>1137</td>
<td>15.2%</td>
</tr>
<tr>
<td>Spanish</td>
<td>1126</td>
<td>15.1%</td>
</tr>
<tr>
<td>French</td>
<td>667</td>
<td>8.9%</td>
</tr>
<tr>
<td>Hindi</td>
<td>657</td>
<td>8.8%</td>
</tr>
<tr>
<td>German</td>
<td>473</td>
<td>6.3%</td>
</tr>
</tbody>
</table>

In total, 50,575 public profiles were collected from LinkedIn, among which 5906 profiles (11.7%) provide a value in the language field. In those profiles, over 70 different languages were provided. However, only six languages are encountered more than 300 times: English, Chinese, Spanish, French, Hindi, German. As all profiles are written in English, English is not considered as a target language. Thus, the remaining five languages are used as target inference languages in the experiments. Detailed statistics about the selected profiles are listed in Table 3-2. Finally, 3566 profiles are selected, each of which specifies the user knowing at least one of the five languages. Two thirds of the profiles are randomly sampled from this set as the training data, and the remaining one third are test data. Note that the positive and negative samples are imbalanced in the collected data where negative samples are much more frequent than positive samples. Therefore, balanced training and testing samples are selected in the experiments.
Like most other social platforms, location is not listed as an attribute of the Education field in LinkedIn profiles. However, this is important evidence for language information inference. This work considers a straightforward way to import the location of each education experience. We defer a deeper study of automatic location identification of study/work experiences using external resources to future work.

LinkedIn has a homepage for many education institutes in the world. These pages have information about the location of the institute. Thus, the location information associated with a person’s study experience in the profiles can be harvested based on the institute names. In total, 3212 different institutes are extracted from the profiles and the locations of 1658 of them are derived (the remaining 1554 institutes are discarded for this research). The location information is attached as an additional attribute of profiles when needed in the experiments.

### 3.4.2. Baseline Methods

This work compares the LSR-FGM with the following four methods of inferring what languages users comprehend based on their social profiles:

1. **RM (Rule-based method):** For each language, this method maintains a full list of countries/regions where this language is used as an official language. These lists are constructed based on a Wikipedia page\(^{20}\) that lists the official language(s) of each country. This method makes an inference decision that a user comprehends a target language only if one of the country/region names in the corresponding list appears in her/his social profile.

2. **RM-L:** This method is almost the same as RM but the input attribute matrix is enhanced with external location information, i.e. the additional location attribute of the institute.

3. **SVM (Support Vector Machine):** This method uses the attribute vector of correlation nodes to train a classification model for each language, and predicts the language information by employing the classification model. The method is implemented with the SVM-light package\(^{21}\).

---


\(^{21}\) [http://svmlight.joachims.org/](http://svmlight.joachims.org/)
4. SVM-L: This method is almost the same as SVM but the input attribute matrix is enhanced with the external location information.

The proposed LSR-FGM itself infers user language information by collectively considering the local textual attributes, user-user social relations and language-language dependency relations. The enhanced attribute matrix is applied on this method.

3.4.3. Evaluation Metrics

Accuracy, precision, recall and F1-score are commonly used as evaluation metrics in various prediction tasks. Following this common practice, we use these metrics to measure the performance of the methods used in this work. To clearly explain these metrics, we will make use of a confusion matrix, as illustrated in Table 3-3. The four entries in the matrix represent four different instances for the prediction of a user's expertise on a language \(l\) by a prediction method. For example, instance TP represents that a user is correctly classified as the user who comprehends language \(l\).

Table 3-3: An example of the confusion matrix for the prediction of user expertise on language \(l\)

<table>
<thead>
<tr>
<th>True Condition</th>
<th>Has Expertise</th>
<th>No Expertise</th>
</tr>
</thead>
<tbody>
<tr>
<td>Predicted Condition</td>
<td>Has Expertise</td>
<td>True Positive (TP)</td>
</tr>
<tr>
<td></td>
<td>No Expertise</td>
<td>False Negative (FN)</td>
</tr>
</tbody>
</table>

Thus, based on the number of occurrences of these four instances over the testing samples, the four measures can be defined as follows for the prediction results of a method on a language \(l\):

\[
\text{Accuracy} = \frac{\#TP + \#TN}{\#TP + \#TN + \#FP + \#FN}
\]

\[
\text{Precision} = \frac{\#TP}{\#TP + \#FP}
\]

\[
\text{Recall} = \frac{\#TP}{\#TP + \#FN}
\]

\[
F1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}
\]
In the following experimental analysis, we use these four metrics to measure the performance of different methods.

### 3.4.4. Performance Comparison and Analysis

Experimental results are listed in Table 3-4. It shows that the LSR-FGM method outperforms all baseline methods overall, on all target languages, and the inclusion of external location information remarkably improves performance.

**Table 3-4 Performance of language inference with different methods on different languages (%)**

<table>
<thead>
<tr>
<th>Language</th>
<th>Metrics</th>
<th>RM</th>
<th>RM-L</th>
<th>SVM</th>
<th>SVM-L</th>
<th>LSR-FGM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chinese</td>
<td>Accuracy</td>
<td>67.79</td>
<td>86.39</td>
<td>83.69</td>
<td>87.74</td>
<td>89.62</td>
</tr>
<tr>
<td></td>
<td>Precision</td>
<td>88.37</td>
<td>93.55</td>
<td>85.92</td>
<td><strong>94.30</strong></td>
<td>85.44</td>
</tr>
<tr>
<td></td>
<td>Recall</td>
<td>40.97</td>
<td>78.17</td>
<td>80.59</td>
<td>80.32</td>
<td><strong>93.24</strong></td>
</tr>
<tr>
<td></td>
<td>F1-score</td>
<td>55.98</td>
<td>85.17</td>
<td>83.17</td>
<td>86.75</td>
<td><strong>89.17</strong></td>
</tr>
<tr>
<td>French</td>
<td>Accuracy</td>
<td>59.77</td>
<td>65.81</td>
<td>63.95</td>
<td>69.30</td>
<td><strong>81.16</strong></td>
</tr>
<tr>
<td></td>
<td>Precision</td>
<td>80.88</td>
<td><strong>85.42</strong></td>
<td>66.85</td>
<td>70.05</td>
<td>75.81</td>
</tr>
<tr>
<td></td>
<td>Recall</td>
<td>25.58</td>
<td>38.14</td>
<td>55.35</td>
<td>67.44</td>
<td><strong>84.90</strong></td>
</tr>
<tr>
<td></td>
<td>F1-score</td>
<td>38.87</td>
<td>52.73</td>
<td>60.56</td>
<td>68.72</td>
<td><strong>80.10</strong></td>
</tr>
<tr>
<td>German</td>
<td>Accuracy</td>
<td>62.16</td>
<td>66.89</td>
<td>66.55</td>
<td>68.58</td>
<td><strong>78.38</strong></td>
</tr>
<tr>
<td></td>
<td>Precision</td>
<td>97.37</td>
<td><strong>98.08</strong></td>
<td>66.23</td>
<td>68.21</td>
<td>74.32</td>
</tr>
<tr>
<td></td>
<td>Recall</td>
<td>25.00</td>
<td>34.46</td>
<td>67.57</td>
<td>69.59</td>
<td><strong>80.88</strong></td>
</tr>
<tr>
<td></td>
<td>F1-score</td>
<td>39.79</td>
<td>51.00</td>
<td>66.89</td>
<td>68.89</td>
<td><strong>77.46</strong></td>
</tr>
<tr>
<td>Hindi</td>
<td>Accuracy</td>
<td>68.70</td>
<td>87.39</td>
<td>82.77</td>
<td>88.24</td>
<td><strong>92.86</strong></td>
</tr>
<tr>
<td></td>
<td>Precision</td>
<td>98.90</td>
<td><strong>99.44</strong></td>
<td>87.86</td>
<td>95.05</td>
<td>91.60</td>
</tr>
<tr>
<td></td>
<td>Recall</td>
<td>37.82</td>
<td>75.21</td>
<td>76.05</td>
<td>80.67</td>
<td><strong>93.97</strong></td>
</tr>
<tr>
<td></td>
<td>F1-score</td>
<td>54.72</td>
<td>85.64</td>
<td>81.53</td>
<td>87.27</td>
<td><strong>92.77</strong></td>
</tr>
<tr>
<td>Spanish</td>
<td>Accuracy</td>
<td>56.22</td>
<td>58.07</td>
<td>74.47</td>
<td>77.12</td>
<td><strong>78.57</strong></td>
</tr>
<tr>
<td></td>
<td>Precision</td>
<td>94.34</td>
<td><strong>95.52</strong></td>
<td>72.84</td>
<td>75.06</td>
<td>82.54</td>
</tr>
<tr>
<td></td>
<td>Recall</td>
<td>13.23</td>
<td>16.93</td>
<td>78.04</td>
<td><strong>81.22</strong></td>
<td>76.47</td>
</tr>
<tr>
<td></td>
<td>F1-score</td>
<td>23.21</td>
<td>28.76</td>
<td>75.35</td>
<td>78.02</td>
<td><strong>79.39</strong></td>
</tr>
</tbody>
</table>

**Performance on different languages**: First, the overall performance of all five methods is better on Chinese and Hindi than on French, German and Spanish. For example, LSR-FGM achieves a 92.77% F1-score on Hindi, while it gets a smaller 80.1% F1-score on the French. This demonstrates that a mix of related languages (referring to French,
German and Spanish) in the target language set increases the difficulty of inference for those languages. Then, we compare the LSR-FGM with other methods on different languages. Table 3-4 shows LSR-FGM outperforms all four baseline methods (in terms of F1-score) but with varying degrees of improvement for different languages. For example, for the target language Chinese, LSR-FGM achieves +2.42% improvement (in terms of F1-score) when compared with SVM-L. By comparison, LSR-FGM significantly outperforms SVM-L by 11.38% (F1-score) on French. This difference in performance between Chinese and French reflects two aspects of information that are also discussed in Section 3.2. First, again it shows the discriminative features of Hindi and Chinese are easier to catch since they share virtually no common characteristics with the other target languages. Second, the relations between languages and profiles significantly contribute to distinguishing these related languages.

**Contribution of additional location attributes:** It is noted that the imported location information plays a crucial role in language inference. Both RM and SVM achieve a much better performance with the enhanced attribute matrix for all languages. For example, SVM-L outperforms SVM by 8.16% (F1-score) on French.

It is observed that RM-L achieves the best precision among all the methods. This is because it tends to predict more negative cases (i.e. fail to find corresponding country names in the profile), which thus hurts recall. For instance, RM-L achieves a very high precision (95.52%) on Spanish, while it gets an extremely low recall (16.93%). It should also be noted that both SVM-L and RM-L achieve high precision on Chinese, with SVM-L slightly outperforming RM-L. This demonstrates that the location-related information provides strong discriminative features for inference of the Chinese language. Therefore, both SVM-L and RM-L are very likely to make a correct inference when the testing user has the location information available for the language Chinese. Although LSR-FGM also includes the additional location information in inference, it typically performs poorer than some of the baseline approaches with regard to precision. This is because its inference decision is made not solely based on the identified textual features but also the defined structural features. This may slightly hurt precision performance for inference, but can greatly help to improve both recall performance and the overall performance (i.e. F1 and accuracy).
Table 3-5: Overall performance of LSR-RGM with different factors (%)

<table>
<thead>
<tr>
<th>Factors</th>
<th>Accuracy</th>
<th>Precision</th>
<th>Recall</th>
<th>F1-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Attributes</td>
<td>80.44</td>
<td>78.22</td>
<td>81.86</td>
<td>80.00</td>
</tr>
<tr>
<td>+Same-experience Rel.</td>
<td>80.78</td>
<td>79.48</td>
<td>81.68</td>
<td>80.54</td>
</tr>
<tr>
<td>+Language Dependency Rel.</td>
<td>84.04</td>
<td>82.59</td>
<td>85.05</td>
<td>83.80</td>
</tr>
<tr>
<td>All</td>
<td>84.52</td>
<td>82.96</td>
<td>85.63</td>
<td>84.27</td>
</tr>
</tbody>
</table>

3.4.5. Factor Contribution Analysis

This subsection examines the contributions of the defined three factors in the LSR-FGM. Table 3-5 gives the overall performance on all target languages of the LSR-FGM, considering the different factors. Specifically, all the relation factors are removed and only the attribute factor is kept, and then each of the relation factors is added into the model separately. The experimental results show that both of the relation factors are useful for inferring what languages the user comprehends. It also indicates that these two relation factors contribute from different perspectives. This is demonstrated by the fact that the LSR-FGM with all three factors outperforms the instances which only consider one relation factor. For example, the same-experience relation could help for those profiles in which only a few study/work experiences are given and not enough discriminative features are available for inferring language information; the language dependency relation could contribute for multilingual users whose profiles only contain enough evidence about certain languages.

3.5. Chapter Summary

This chapter of PhD work studies the novel problem of inferring what languages a user comprehends based on their static social profiles and aims to exploit this limited information about cold-start users in SNSs to infer their language expertise. It precisely defines the problem and proposes a language and social relation-based factor graph model. This model collectively considers three factors in the inference process: textual attributes of the social profile; dependency relations between target languages; and social relations between users. The experiment on a real-world large-scale dataset shows the success of the proposed model in inferring user language information using social profiles, and demonstrate that each one of the three factors makes a stand-alone, complementary contribution in the model. In addition, this work proposes to import
information reflecting users’ language information from external resources in order to help inference, which is shown to be effective in the experiments.

4. Topical Expertise Inference on Twitter: Integrating Sentiment and Topic Relatedness

4.1. Introduction

In Chapter 3, this PhD research attempted to exploit static profiles of cold start users in SNSs to infer their language expertise information. As discussed in Chapters 1 and 2, existing studies aimed at expertise discovery in SNSs primarily focus on the exploitation of users’ rich activities, i.e. targeting non-cold start users. This chapter of the PhD thesis focuses on the most popular micro-blogging site, Twitter, and aims to infer a Twitter user's topical expertise based on the content (i.e. tweets) she/he has posted (the use of further user features for user expertise inference will be discussed in Chapter 5).

As discussed in Section 2.1, in the state of the art, there have been a number of attempts made [Wa12, Bo13, Gh12] to model the problem of user expertise inference on Twitter. These can be broadly classified into two categories.

The first, is to simplify the problem as a standard information retrieval problem, i.e. given an expertise topic (query terms) and a list of Twitter users (documents built from the tweets of each user), the objective is to compute the degree of relevance between the expertise topic and each user. This method can obtain a user ranking with respect to an expertise topic, with the expectation being that more knowledgeable people on this topic are ranked highest. But this approach suffers from the problem of not being able to clearly discriminate knowledgeable users from others. In social media environments, users can publish content on any topic, so the frequent mentions of terms about a topic in a user's authored content does not necessarily indicate that this user has a good level of knowledge of that topic. Thus, understanding how to effectively utilise the user’s posted content in the inference of their expertise is an important challenge.

The second commonly used method is to model the problem as a supervised classification problem [Wa12, Gh12]. In other words, predicting what expertise topics a user has knowledge of relies on features extracted from tweets posted by the user. This method often suffers from a shortage of labeled data, because Twitter does not have an expertise attribute in user profiles, and even if it did, users typically do not want to go to the effort of maintaining such information. Existing studies that apply this method mainly rely on manual labeling for obtaining training data [Po13], which is expensive and time-
consuming. This presents a significant obstacle to the development of research in this area. Similarly, this approach also faces the challenge of how to effectively use the user posted tweets.

To address these challenges, the research described in this chapter proposes exploiting two types of prior knowledge for user expertise inference on Twitter: (i) the sentiment intensity of tweets and (ii) relations between expertise topics. In addition, this work builds a large-scale dataset for the study by using cross-SNS resources.

(1) Sentiment Intensity of tweets: There is uncertainty around the incentive that triggers a user's posts or tweets. For example, a user may tweet to support a friend in spreading information, in this case the user does not require any knowledge on that topic. Thus, a user does not necessarily have good knowledge of the expertise topics contained in every tweet. This work proposes to analyse the sentiment intensity of the tweet and use this to evaluate the importance of a tweet in inferring a user's topical expertise. There is an intuition that if a person can express their opinion forcefully on a topic, either positive or negative, they are perceived as knowledgeable in this domain. This intuition is also validated by experimentation of this research. So, it is reasonable to assume in user expertise inference that tweets expressing intense sentiment are more likely to reflect the topic expertise of a user.

(2) Relations between expertise topics: Previous studies take expertise inference on each topic as a separate task. This goes against the inherent characteristics of a person's knowledge distribution. The number of expertise topics a user has knowledge of is limited and there are intrinsic relations between the topics of expertise of an individual. For example, a user who knows the Java programming language is more likely to know the C programming language than a user who knows no programming languages. This prior knowledge could reflect the probability that a user knows other expertise topics based on a topic the user has knowledge of and prevent the model from unrealistically estimating the knowledge coverage of a person. Thus, this work proposes to simultaneously infer a user's expertise in multiple topics and exploit the prior relations between topics to model the expertise inference problem.

(3) Dataset construction using cross-SNS resources. To overcome the problem of the shortage of training data, this work proposes utilising the account connection between the popular Community Question Answering (CQA) site Quora, and Twitter to automatically
build a large-scale dataset. In CQA sites, users actively use their knowledge to help question askers answer their questions, which effectively demonstrates their expertise. In Quora, users are encouraged to explicitly provide their knowledge areas for accurate and efficient matching between questions and answerers. Besides, it is observed that many Quora users also have SNS accounts like Twitter and LinkedIn. A big proportion of them directly provide links to their SNS accounts in their profiles for the convenience of question askers gaining a better understanding of the credibility of their answers [Zh14]. Thus, this naturally raises the idea of using the content of an individual’s Twitter account in combination with the expertise topics they specified in Quora for dataset construction. This work utilises the knowledge of user-generated expertise topics and the CQA-SNS (i.e. Quora-Twitter) connections of users to build a large-scale training dataset for expertise inference.

Finally, this chapter proposes a Sentiment-weighted and Topic Relation-regularized Learning (SeTRL) model to infer the topical expertise of Twitter users based on their tweets. Experiments are conducted on a large-scale dataset with over 10,000 Twitter users and 149 expertise topics. The experimental results demonstrate that sentiment analysis and the topic relation regularizer can significantly improve the prediction accuracy of topical expertise of Twitter users.

The rest of the chapter is organized as follows. Section 4.2 formally defines the problem of user expertise inference on Twitter and then details the proposed SeTRL model; Section 4.3 describes the construction of our experimentation datasets; Section 4.4 gives the experimental results and analysis; Section 4.5 concludes this chapter.

4.2. Sentiment-Weighted and Topic Relation-Regularized Learning for User Expertise Inference

This section first formally defines the problem of expertise inference of Twitter users and then introduces the scheme that uses tweet sentiment analysis to assess the importance of different tweets in expertise inference. Finally, it details the construction of the SeTRL model and proposes a method to learn the model.

4.2.1. Problem Definition

This sub-section formally defines the problem of expertise inference of Twitter users. The definition of notations can be found in Table 4-1. The input of the problem is $N$ Twitter
users, and \( T \) expertise topics. Each user \( i \) for the topic \( t \) can be represented by a \( k \)-dimension feature vector \( \mathbf{x}_i^{(t)} \). This feature vector is defined from the tweets of a user that can reflect the user's expertise on a certain topic. It would be reasonable to design a dedicated feature space for the inference of each expertise topic, however, the focus of this research is to study the influence of sentiment analysis and topic relatedness on expertise inference of Twitter users, rather than feature engineering for hundreds of topics. Therefore, this work adopts the general unigram features from the tweets as the feature space for the inference of each topic. Then, we have a \( N \)-by-\( K \) feature matrix \( X \) in which each row corresponds to a user, each column a feature of the user, and an entry \( x_{ij} \) denotes the feature value of the \( j^{th} \) feature of user \( i \). A corresponding \( T \)-by-\( N \) label matrix \( Y \) is defined to represent the expertise of users on the \( T \) topics, in which each column corresponds to a user, each row a topic, and \( y_{ti} \in \{+1, -1\} \) denotes the expertise of user \( i \) on topic \( t \). The objective of this work is to learn a predictive model for every expertise topic that can effectively infer a user's expertise on the topic.

### Table 4-1: Definition of Notations

<table>
<thead>
<tr>
<th>Notation</th>
<th>Notation Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>( N )</td>
<td>Number of users</td>
</tr>
<tr>
<td>( T )</td>
<td>Number of expertise topics</td>
</tr>
<tr>
<td>( K )</td>
<td>Number of user features</td>
</tr>
<tr>
<td>( x_i )</td>
<td>Feature vector of the ( i^{th} ) user</td>
</tr>
<tr>
<td>( y_{ti} \in {+1, -1} )</td>
<td>Has or doesn’t have expertise</td>
</tr>
<tr>
<td>( G, E )</td>
<td>Topic relation graph</td>
</tr>
<tr>
<td>( w_t )</td>
<td>Model coefficients of topic ( t )</td>
</tr>
<tr>
<td>( \alpha, \beta )</td>
<td>Regularization parameters</td>
</tr>
<tr>
<td>( \text{sentiment_threshold} )</td>
<td>Threshold for tweet selection</td>
</tr>
</tbody>
</table>

#### 4.2.2. Feature Weighting with Tweet Sentiment Analysis

The first issue is to weight the features of every user for expertise inference. As discussed in Section 4.1, it is the contention of this research that a user's opinion on a topic can more effectively reflect the user's expertise on that topic rather than frequent mentions of terms related to the topic. This intuition is validated in the experiments conducted in this research. Thus, it is proposed to weight a user's feature vector based on the sentiment intensity of their posted tweets. Due to the lack of a training dataset for learning-based methods and its superior performance for social texts like tweets, this work adopts a
lexicon-based method called VADER [Hu14]. A detailed discussion about the selection of sentiment analysis methods is presented in Section 2.4.2.

In addition, we use a sentiment threshold to further select which tweets will be considered in the weighting scheme. This is based on two points of consideration. Firstly, there may exist errors in tweet sentiment assessment, especially for those boundary cases. It means tweets with neutral sentiment orientation are more likely to be mistakenly considered as low-intensity sentiment tweets, which could harm the user expertise prediction. Secondly, only considering tweets with high-intensity sentiment would overly restrict the effect of the tweet sentiment in expertise inference. Thus, we propose a maximum-value scheme to calculate the weight $x_{ij}$ of the $j^{th}$ feature of the user $i$ based on the sentiment intensity of that individual’s tweets:

\begin{algorithm}
\textbf{Algorithm 4-1: Feature Weighting with Sentiment Analysis}
\textbf{Input}: tweets of user $i$, sentiment\_threshold
\textbf{Output}: $x_i$
1. for the $j^{th}$ feature of user $i$:
2. \hspace{0.5cm} for tweet $tw$ that contains the $j^{th}$ feature in tweets of user $i$:
3. \hspace{1cm} if $\text{sentscore}(tw)$ > sentiment\_threshold:
4. \hspace{1.5cm} $x_{ij} = \max(x_{ij}, |\text{sentscore}(tw)|)$
5. \hspace{1cm} end if
6. \hspace{0.5cm} end for
7. end for
\end{algorithm}

In the Algorithm 4-1, $\text{sentscore}(tw)$ is the sentiment score of the tweet $tw$ generated by VADER. We compare our scheme with several other alternative weighting schemes and analyse their experimental results in Section 4.4.

4.2.3. Expertise Inference with Topic Relation Regularization

This section details the construction of the complete Sentiment-weighted and Topic Relation-regularized Learning (SeTRL) model. To take the topic relatedness into consideration in the process of expertise inference, it first needs to construct a model that can simultaneously infer a user’s expertise on multiple topics. As presented in the previous section, the problem of expertise inference is formulated as a classification problem.
Thus, for each topic $t$, we can easily learn a separate predictive model based on linear regression, and predict the label of user $i$ on topic $t$ as follows:

$$y_{ti} = f_t(x_i) = x_i w_t$$

(4.1)

where $x_i$ is the feature vector of user $i$; $w_t$ is model parameter vector that represents the linear mapping function $f_t$ of topic $t$. By using linear regression, a base model for jointly learning the expertise of users on multiple expertise topics can be formulated as learning an optimal solution $W = [w_1, w_2, \ldots, w_T]$ by solving the following minimization problem:

$$\min_W \sum_{i=1}^{N} \sum_{t=1}^{T} \ell(y_{ti}, f_t(x_i)) + \beta ||W||_1$$

(4.2)

where $||W||_1$ denotes the $l_1$ norm of matrix $W$ which is used to address the model sparsity problem and stabilise Eq. (2) [Ti96] and $\beta$ is the regularization parameter to control the sparsity; $\ell$ is the least square loss function for regression:

$$\ell(y_{ti}, f_t(x_i)) = \frac{1}{2} (y_{ti} - x_i w_t)^2$$

(4.3)

which is equivalent to a linear discriminant analysis for binary classification [Fr01a] and has been widely applied in various classification algorithms [Ri03].

Thus, a base model has been built that can simultaneously learn the predictive models of all expertise topics. However, this optimization model simply assumes that each inference task selects its own relevant features individually, which might be unrealistic in practice. For example, the inference of expertise topics "data science" and "machine learning" tend to share a common set of relevant features but these features are less likely to be useful for the topic "gardening". As also discussed in Section 4.1, expertise topics of a person could be related to each other, e.g. C language and Java programming language. This relatedness of topics could reflect the fact that they share common discriminant features. This observation propels us to characterise the relatedness of topics by an undirected graph $G$ with $E$ edges and encode this structure information in the base model by using a Tikhonov regularizer [Go99]:

$$\sum_{e=1}^{E} ||w_{e(1)} - w_{e(2)}||_2^2$$

(4.4)
where $w_{e(1)}$ is the model parameters corresponding to a topic of edge $e$, where an edge connects two related topics. This regularizer can penalize the difference of two inference tasks that have an edge between them, through which we can incorporate the topic relatedness in the model leaning process. Two specific methods of constructing the topic relation graph $G$ are introduced in Section 4.2.5.

Now, the SeTRL model can be constructed based on the above formulation. By combining Eqs. (4.2)-(4.4), we have the following optimization problem:

$$
\min_W \sum_{i=1}^{N} \sum_{t=1}^{T} \frac{1}{2} (y_{ti} - x_i w_t)^2 + \alpha \sum_{e=1}^{E} \|w_{e(1)} - w_{e(2)}\|^2_2 + \beta \|W\|_1
$$

(4.5)

where $\alpha$ is the regularization parameter to control the contribution of the topic relation information.

### 4.2.4. Model Learning and Prediction

The next issue to examine is how to learn the SeTRL model, i.e. to estimate a parameter configuration of $W$ that minimizes the objective function Eq. (4.5). Because the $l_1$ norm term in Eq. (4.5) (i.e. $\|W\|_1$) is not differentiable, the gradient descent method is not suitable for the model learning. In this work, we use the Accelerated Proximal Gradient (APG) method [Ji09] to solve the optimization problem. Different from the traditional gradient method that uses the latest point in every iteration, APG takes a linear combination of the previous two points as the search point to achieve high convergence speed.

Specifically, we first decompose the objective function Eq. (4.5) to a non-smooth part:

$$
g(W) = \beta \|W\|_1
$$

(4.6)

and a smooth convex part:

$$
h(W) = \sum_{i=1}^{N} \sum_{t=1}^{T} \frac{1}{2} (y_{ti} - x_i w_t)^2 + \alpha \sum_{e=1}^{E} \|w_{e(1)} - w_{e(2)}\|^2_2
$$

(4.7)

Then, the generalised gradient updating step can be approximately expressed by:

$$
R_{s}(W, W_{(m)}) = h(W_{(m)}) + \left\langle W - W_{(m)}, \nabla h(W_{(m)}) \right\rangle
$$

$$
+ \frac{\beta}{2} \|W - W_{(m)}\|_1^2 + g(W)
$$

(4.8)
\[ r_s(W_{(m)}) = \arg \min_{W} R_s(W, W_{(m)}) \]  

(4.9)

where \( \| \cdot \|_F \) denotes the Frobenius norm; \( \nabla h(W_{(m)}) \) is the gradient of \( h(W) \) at the point \( W_{(m)} \) of the \( m \)th iteration and \( \langle A, B \rangle = \text{tr}(A^T B) \) denotes the inter product of matrixes; \( s \) is the iteration step size. After rewriting Eq. (4.9), we can have:

\[ r_s(W_{(m)}) = \arg \min_{W} \frac{1}{2} \| W - V_{(m)} \|_2^2 + \frac{1}{s} g(W) \]  

(4.10)

where \( V_{(m)} = W_{(m)} - \frac{1}{s} \nabla h(W_{(m)}) \) and \( s \) can be determined by using line search.

The key aspect of the APG method is how to efficiently update the parameter matrix \( W \) in each iteration. Finally, we adopt the strategy proposed in [Ch09] to further optimize the problem of Eq.(4.10). The gradient update rule of Eq.(4.10) is decomposed to \( K \) sub-problems, for which the analytical solutions can be easily obtained. In addition, instead of performing gradient descent based on \( W_{(m)} \), the search point is computed as:

\[ U_{(m)} = W_{(m)} + \sigma_m(W_{(m)} - W_{(m-1)}) \]  

(4.11)

where \( \sigma_m = (1 - \lambda_{m-1}) \lambda_m / \lambda_{m-1} \) and \( \lambda_m = 2/(m+3) \). In Algorithm 4-2, we summarize the detailed optimization procedure.

---

**Algorithm 4-2 APG Algorithm for Learning ScTRL**

**Initialization:** \( s_0 > 0, \eta > 1, W(\theta) \in \mathbb{R}^{K \times T}, U(\theta) = W(\theta) \) and \( \lambda_0 = 1, m = 0 \)

**Output:** \( W_{(m)} \)

1. for \( m = 0, 1, 2, \ldots \) until convergence of \( W_{(m)} \):
   2. Set \( s = s_m \);
   3. while \( (g(r_s(U_{(m)})) + h(r_s(U_{(m)})) > R_s(r_s(U_{(m)}), U_{(m)})) \):
      4. Set \( s = \eta s \);
   5. end while
   6. Set \( s_{m+1} = s \);
   7. \( W_{(m+1)} = \arg \min_W R_{s_{m+1}}(W, U_{(m)}) \);
   8. \( \lambda_{m+1} = 2/(m+3) \);
   9. Compute \( U_{(m+1)} \) using Eq.(4.11);
10. end for
4.2.5. Construction of Topic Relation Structure

To learn the SeTRL model, the last issue is how to construct the topic relation graph $G$ that can effectively reflect the connections between the expertise topics of a user. Due to the diversity of incentives that drive people to acquire knowledge about something, it is challenging to accurately capture the relatedness of any two expertise topics. In this research, we utilise two types of prior knowledge to build $G$.

(1) **Internal source.** Intuitively, if many users appear to have knowledge of the same two topics, it likely indicates that the two topics are related. Driven by this consideration, we propose to build the relatedness of any two topics based on their co-occurrence in users' Quora profiles in our dataset. Note that a user specifies their expertise topics in their Quora profile, which is detailed in Section 4.3. Specifically, two topics are considered as related, i.e. connected with an edge $e$ in $G$, if their number of co-occurrence is larger than the average co-occurrence over all topic pairs in all collected Quora profiles.

(2) **External source.** If two expertise topics are close in their meaning, such as "health and nutrition" and "nutritional supplements", this could also suggest a likelihood that a user has expertise on both of the topics. This inspires us to build the relatedness of topics based on the similarity of topic meanings. By employing the Wu-Palmer Similarity \cite{Wu94} that assesses the similarity of two word senses based on the word hierarchical structure in the WordNet ontology, we compute the similarity of two topics as follows:

$$
\text{Similarity}(t_1, t_2) = \sum_{w_{d_1} \in t_1} \sum_{w_{d_2} \in t_2} \frac{Wu - Palmer(w_{d_1}, w_{d_2})}{\|w_{d_1}, w_{d_2}\|}
$$

(4.12)

where $w_{d_i}$ is a word in the topic name of $t_i$; $\|(w_{d_1}, w_{d_2})\|$ is the number of word pairs from the two topic names. Finally, two topics are connected as related topics if their similarity score is larger than the average similarity score over all topic pairs.

4.3. Dataset Construction

To construct a benchmark dataset, we needed to identify the expertise topics of a large number of Twitter users. Manual annotation of such a dataset would be unrealistic, due to the time-consuming and painstaking work required. Therefore, as mentioned in Section 4.1, this work proposes to automatically annotate the expertise topics of a Twitter user through the popular question and answering service Quora. This is based on the fact that a large proportion of Quora users explicitly provide both their expertise topics and Twitter
account information in their profiles. Figure 4-1 shows an example of a Quora profile, in which two red boxes mark the Twitter account information and expertise topics of the Quora user respectively.

![Quora Profile Example](image)

**Figure 4-1: An example of a Quora Profile**

A crawler was first employed to harvest user profiles in Quora, from which only those profiles which contain both the user’s Twitter account and expertise information are extracted. For privacy protection, the names of the profiles are not collected. Then, tweets of the harvested Twitter users are collected using the official Twitter API\(^{22}\), which applies an upper limit of 3200 tweets for each Twitter account. For each Twitter user, the expertise topics specified in their Quora profile are taken as the ground truth of their expertise information. To obtain a representative set of expertise topics and Twitter users, we filter out topics that occurred less than 50 times over all the users and users that posted

\(^{22}\)https://dev.twitter.com/
less than 100 tweets. Finally, we obtain a dataset with 10,856 Twitter users and 149 expertise topics, in which every user has knowledge of at least one topic from the 149 topics. Table 4-2 provides descriptive statistics of the dataset. The 149 topics cover a wide range of knowledge areas from “software engineering”, to “atheism” and “religion”. There are also topics that are closely related, such as “data science” and “machine learning”. In the experiment, it is noted that for every topic the negative samples are much more frequent than positive samples. Therefore, balanced positive and negative samples of each topic are randomly chosen from the dataset, two thirds of which are used as a training set in the experiment. The remaining one third is the test set for the topic.

<table>
<thead>
<tr>
<th>Item</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Users</td>
<td>10,856</td>
</tr>
<tr>
<td>Topics</td>
<td>149</td>
</tr>
<tr>
<td>Tweets of all users</td>
<td>7,202,033</td>
</tr>
<tr>
<td>Avg. tweets per user</td>
<td>663</td>
</tr>
<tr>
<td>Avg. topics per user</td>
<td>2.4</td>
</tr>
</tbody>
</table>

### 4.4. Experiment and Results

#### 4.4.1. Baseline Methods and Evaluation Metrics

This work compares SeTRL with the following methods of inferring topical expertise of Twitter users based on their tweets:

1. **SVM**: This method uses the binary unigram features to build the feature vectors of users and then the Support Vector Machine (linear kernel) is employed to train a classifier for each topic. The method is implemented with the SVM-light package.

2. **SVM-Bi**: This method is almost the same as SVM but the user feature vector is built using both the binary unigram and bigram features.

3. **SVM-Sen**: This method is almost the same as SVM but the user feature matrix is weighted using our proposed tweet sentiment based weighting scheme.

4. **SVM-AvgSen**: This method is a derivation of SVM-Sen. A feature is weighted by using the averaged absolute sentiment intensities of tweets that contain this feature.
5. SVM-TFIDF: This method is another derivation of SVM-Sen. It uses the TF-IDF [Ai03] to assign weights to the user features where all the tweets of a user are considered as a document.

6. TRL: This method is similar as SeTRL but it adopts the binary feature matrix.

Two methods of constructing a topic relation graph are applied to both TRL and SeTRL: co-occurrence based method and Wordnet similarity based method. They are denoted as TRL-Co, TRL-WN and SeTRL-Co, SeTRL-WN respectively in the following subsection.

Similar to the last chapter, in the following experiments, the four metrics: accuracy, precision, recall and F1-score are used to examine the overall performance of different methods for parameter configuration and method comparison, i.e. the performance of different methods on all the expertise topics. In addition, the performance of the methods on specific topics is also presented and discussed.

4.4.2. Parameter Configuration

Sentiment threshold selection: As discussed in Section 4.2.2, we need to select an appropriate parameter sentiment_threshold to determine which tweets will be used in the proposed weighting scheme. Therefore, we observe the performance of SVM-Sen with different threshold values for the optimal threshold selection. Figure 4-2 and Figure 4-3 show the overall F1 and accuracy performance of SVM-Sen and SVM respectively with the varying threshold between [0, 1], i.e. when \(|\text{tweet sentiment intensity}| > \text{sentiment_threshold}\) in algorithm 4-1. They show that SVM-Sen consistently outperforms SVM under any threshold configuration, however this reaches the best performance at a threshold of 0.4. This result reflects two aspects of information that are also discussed previously. First, the proposed tweet sentiment based weighting scheme can effectively improve the performance of expertise inference of Twitter users. Second, either incorporating too many low-intensity sentiment tweets or overly restricting the selection of sentiment-bearing tweets in our weighting scheme can harm the inference accuracy. In the following experiments, we set the threshold as 0.4 for SVM-Sen, SeTRL-Co and SeTRL-WN.

Parameters α, β configuration: In the experiments, we perform the standard 5-fold cross-validation on the training dataset to select the regularization parameter α, β. The method of grid search in range of $10^{-4}$ to $10^3$ with the power increased by 1 is used to tune the parameters. For example, Figure 4-4 and Figure 4-5 show the overall performance of
SeTRL-Co with different configurations of $\alpha$ and $\beta$ respectively (the other parameter is set as the optimal value). They suggest that the optimal parameter configuration is $\alpha = 100$, $\beta=0.001$. Thus, this parameter configuration is selected to learn a final SeTRL-Co model from the entire training set and apply the model to the testing dataset for prediction. This procedure is also applied to TRL-Co, TRL-WN and SeTRL-WN.

Figure 4-2: F1 performance of SVM-Sen and SVM with the different sentiment threshold

Figure 4-3: Accuracy performance of SVM-Sen and SVM with the different sentiment threshold
4.4.3. Performance Comparison and Analysis

**Overall performance:** Table 4-3 presents the overall performance of different methods. In section 4.4.2, it is shown that our proposed tweet sentiment based weighting scheme can effectively improve the expertise inference of Twitter users. As shown in Table 4-3, the significant improvements of SeTRL-Co, SeTRL-WN over TRL-Co, TRL-WN
respectively, again demonstrate the effectiveness of our scheme. In addition, we compare it with another two weighting schemes, which correspond to the two methods SVM-AvgSen and SVM-TFIDF (described in section 4.4.1). Table 4-3 shows that SVM-Sen significantly outperforms the two methods on all four metrics. In particular, the TF-IDF algorithm that is commonly adopted in the field of information retrieval shows inferior performance in the task of expertise inference, compared with other tweet sentiment based methods. This result supports our two previous assumptions. First, that frequent mentions of terms about a topic in a user's authored social content does not necessarily indicate that this user has a good level of knowledge of that topic. Second, strong opinion, either negative or positive, expressed by a user on a topic likely suggests this user has knowledge of this topic to some extent.

Table 4-3: Overall performance of expertise inference with different methods (%)

<table>
<thead>
<tr>
<th>Method</th>
<th>Recall</th>
<th>Precision</th>
<th>F1</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM</td>
<td>69.09</td>
<td>66.55</td>
<td>66.06</td>
<td>65.79</td>
</tr>
<tr>
<td>SVM-Bi</td>
<td>70.14</td>
<td>67.55</td>
<td>66.32</td>
<td>66.02</td>
</tr>
<tr>
<td>SVM-Sen</td>
<td>72.10</td>
<td>70.88</td>
<td>69.76</td>
<td>69.75</td>
</tr>
<tr>
<td>SVM-AvgSen</td>
<td>70.24</td>
<td>70.20</td>
<td>69.98</td>
<td>66.04</td>
</tr>
<tr>
<td>SVM-TFIDF</td>
<td>55.13</td>
<td>70.03</td>
<td>60.16</td>
<td>66.03</td>
</tr>
<tr>
<td>TRL-WN</td>
<td>79.62</td>
<td>76.67</td>
<td>77.61</td>
<td>77.17</td>
</tr>
<tr>
<td>TRL-Co</td>
<td>79.27</td>
<td>77.73</td>
<td>78.00</td>
<td>77.91</td>
</tr>
<tr>
<td>SeTRL-WN</td>
<td>81.82</td>
<td>76.86</td>
<td>78.71</td>
<td>77.96</td>
</tr>
<tr>
<td>SeTRL-Co</td>
<td><strong>82.49</strong></td>
<td><strong>78.76</strong></td>
<td><strong>80.08</strong></td>
<td><strong>79.65</strong></td>
</tr>
</tbody>
</table>

**Contribution of topic relatedness:** We now examine the contribution of the prior knowledge of topic relatedness in our model. As shown in the Table 4-3, SeTRL-Co achieves a 10.32% (F1 score) improvement, compared with SVM-Sen. Similarly, SeTRL-WN also significantly outperforms SVM-Sen by 8.95% in terms of the F1 score. This demonstrates that the inclusion of topic relatedness in the inference model can remarkably improve inference performance. It is also noted that SeTRL-Co and TRL-Co slightly outperform SeTRL-WN and TRL-WN respectively in terms of all four measures. This indicates that the co-occurrence of topics that a user has knowledge of can more effectively reflect the relatedness of expertise topics, compared with using the semantic relatedness in Wordnet.
Performance on topics with different sizes of training data: This subsection examines the performance of our model on topics with different sizes of training data. The 149 topics are first classified to 3 different groups according to the size of their training samples: size < 100 (59 topics), 100 ≤ size ≤ 300 (54 topics) and size > 300 (36 topics). We then obtain the overall F1 performance of SVM, SVM-Sen and SeTRL-Co on the three groups of topics respectively, as shown in Figure 4-6. In addition, we also take a specific example (topic) from each of the groups, topic names of which are "Mobile Marketing" (training data size = 80), "Public Relations" (training data size = 176) and "Machine Learning" (training data size = 343) respectively. Figure 4-7 presents the F1 performance of SVM, SVM-Sen and SeTRL-Co on the three topics. They show that the larger the size of the training data, the better performance SVM and SVM-Sen achieve. However, SeTRL-Co has a similar F1 performance on all three groups of topics. This indicates that taking topic relatedness into consideration in the model can be a great help to the inference of topics with limited training samples. This is because they can directly benefit from the case where there is not enough evidence obtainable from a user's tweets for inferring the user's expertise on a topic, but the user's other expertise topics are available.
Feature space selection: Finally, we examine the effect of feature space selection. It shows that the SVM and SVM-Bi achieve very similar performance. This means that increasing the size of the feature space by several times does not significantly improve the prediction accuracy, but significantly increases the computational complexity. This explains why we only adopt the fundamental unigram feature space in this work.

4.5. Chapter Summary

This chapter of the PhD thesis targets the problem of expertise modelling of non-cold start users in SNSs, and, specifically, studies the problem of inferring what expertise topics a Twitter user has knowledge of based on their tweets. A sentiment-weighted and topic relation-regularized learning model is proposed to model this inference problem. This model uses tweet sentiment analysis to assess the importance of each tweet in expertise inference. It also considers the relatedness between expertise topics as an important aspect of prior knowledge to optimize the inference model. Additionally, this work explores the account connection between the community question answering site Quora and Twitter to automatically harvest training data for learning inference models on various expertise topics. Experiments conducted on a large-scale dataset show that our proposed model outperforms the baseline approaches, and demonstrates that the tweet
sentiment based weighting scheme as well as the topic relation regularizer make a stand-alone contribution in the model.

The experimental work described in this chapter was published in the following paper:

5. Topical Expertise Inference on Twitter: Combining Multiple Types of User Activity

5.1. Introduction

The previous chapter focused on the exploitation of Twitter users’ tweets to infer their topical expertise. However, apart from posting tweets, Twitter users can interact with other users through various means. For example, they can mention another user in a tweet; a user can “follow” other users to be updated on their recent activities; a user can “heart” or “retweet” another user’s tweets. These activities expose a user to multiple types of data on Twitter\(^{23}\). In addition to the work of the last chapter, that explored the relation between a user’s tweets and his/her topical expertise, there are other studies [Su11, We10, Gh12] that have also observed the connections between the user’s other activities and their expertise. For example, the short bio provided by the user on their profile was used to identify topic experts on the “who to follow” service of Twitter [Su11]; Previous research [We10] verified the existence of homophily among Twitter user “following” relationships, i.e. Twitter users tend to follow other users with common topical beliefs or interests. Therefore, the “following” relationships between users were exploited to discover influential users on different topics.

However, these studies tend to focus on the exploitation of a certain type of user data and the potential relationship between this data and the user’s expertise information. Although shown to be effective in inferring a user’s expertise information, these approaches ignore the fact that many Twitter users may not have certain types of data. For example, on Twitter it is reported that approx. 44% of all registered users have never posted a tweet, and most tweets are generated by a small proportion of the user population\(^{24}\); Statistical analysis from about 10% of the entire Twitter population shows that on average, each user is included in less than one Twitter list [Ki10], which is used as prime evidence to infer the user’s topics of expertise in [Gh12]. Moreover, statistics from the Twitter dataset collected as part of this research show that 24.42% of Twitter users follow less than 100 people and 24.18% have less than 100 followers. Therefore, approaches that rely on a

\(^{23}\) In this work, different types of user data refer to, for example, tweets from a user, lists a user is part of, and the list of accounts a user follows and is followed by.

single type of user data will fail when the user does not have a significant volume, or any, of this data available.

To address this issue, this research proposes a learning-based model that tries to infer a user’s expertise information by jointly exploiting multiple types of data associated with the user on Twitter, such as the user’s posted tweets and the followers of the user. The proposed model is named the multi-Data and Topic relatedness Combined (D^oTCom) learning model in this thesis. It aims to make the most of various data associated with the user and ensure the inference effectiveness regardless of the availability of some types of user data. Meanwhile, D^oTCom considers the consistency of different types of user data in the process of inference, which means that the expertise information reflected by different types of user data should be similar. Through regularization, D^oTCom tries to penalize the differences among the inference results from different types of user data.

Experiments are conducted on the same Twitter dataset constructed in the last chapter which has over 10,000 users and 149 expertise topics. Four types of data associated with the user are considered in the experiments, namely: tweets, friends, followers and lists. Experiments first demonstrated that using each type of user data alone can effectively infer a user’s expertise but with varying degree of effectiveness. Experimental results then showed that our proposed model, which combines all the different types of user data, outperforms the alternative inference methods, which use only one type of data, or the combination of fewer types of user data.

In summary, the contributions of this research are twofold:

(1) This work proposes a learning model, D^oTCom, that infers a user’s topical expertise based on multiple types of data associated with him/her on Twitter. It can deliver effective inference once there are some types of user data available, and the model can also further improve the inference accuracy by making use of multiple types of user data if available.

(2) Experiments conducted on a real-world Twitter dataset demonstrate that each of the four types of user data (tweets, friends, followers and lists) is effective for user expertise inference, and show that our D^oTCom learning model which combines multiple types of user data outperforms a number of baseline approaches.

25 “Friends” in the chapter refers to the same meaning as “Followees” used in other chapters, i.e. other users the users is following on Twitter.
The rest of this chapter is organized as follows. Section 5.2 defines the problem of user expertise inference from different types of user data on Twitter and then details our proposed D^2TCom learning model to address this problem; Section 5.3 describes the construction of the experimentation dataset and model input features; Section 5.4 gives the experimental results and analysis; Section 5.5 concludes this chapter of work.

5.2. User Expertise Inference Using Multiple Types of Data on Twitter

This section formally defines the problem of user expertise inference based upon multiple types of user-related Twitter data and then details the construction of our model for addressing the problem.

5.2.1. Problem Definition

In the last chapter, we formally defined the problem of user expertise inference on Twitter, with the focus on the utilisation of the user’s tweets. In comparison, this chapter of work aims to better model the problem by making use of multiple types of user data, which includes the user’s tweets, but is not limited to that. In particular, this work considers four different types of user data for inferring a user’s expertise: (1) Tweets: the textual content posted by the user; (2) Friends: other Twitter users that the user is following; (3) Followers: other Twitter users who are following this user; (4) Lists: Twitter lists that include this user.

Formally, the input is $T$ expertise topics and $S$ feature matrices ($S$ equals to 4 in this work) of $N$ Twitter users: $X_1, X_2, \ldots, X_S$ where $X_s$ is a $N$-by-$K_s$ matrix defined from the social data associated with the user through the $s^{th}$ relationship, i.e. the $s^{th}$ type of user data on Twitter, and $K_s$ is the number of features defined from this data source. An entry $x_{sij}$ of $X_s$ denotes the feature value of the $j^{th}$ feature of user $i$ from the $s^{th}$ data source. The output of the question is the same as the inference problem defined in the last chapter, i.e. a $T$-by-$N$ label matrix $Y$ which represents the expertise of users on the $T$ topics. An entry $y_{ti}$ of $Y$ is a binary value {+1, -1} which denotes the expertise of user $i$ on topic $t$. The objective of this work, is to learn a predictive model for every expertise topic that can effectively infer a user’s expertise on the topic, based on multiple types of data associated with the user on Twitter. Table 5-1 gives the definition of the notations used in this chapter. Please note that the same notations are used from Chapter 4 if they represent the same concepts.
### Table 5-1: Definition of Notations

<table>
<thead>
<tr>
<th>Notation</th>
<th>Notation Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Data</strong></td>
<td></td>
</tr>
<tr>
<td>( N )</td>
<td>Number of users</td>
</tr>
<tr>
<td>( T )</td>
<td>Number of expertise topics</td>
</tr>
<tr>
<td>( S )</td>
<td>Number of types of user data</td>
</tr>
<tr>
<td>( K_s )</td>
<td>Number of user features for the ( s^{th} ) type of user data</td>
</tr>
<tr>
<td>( x_{si} )</td>
<td>Feature vector of the ( i^{th} ) user from the ( s^{th} ) type of user data</td>
</tr>
<tr>
<td>( y_{si} \in {+1, -1} )</td>
<td>Has or doesn’t have expertise</td>
</tr>
<tr>
<td>( G, E )</td>
<td>Topic relation graph</td>
</tr>
<tr>
<td><strong>Model</strong></td>
<td></td>
</tr>
<tr>
<td>( w_t )</td>
<td>Model coefficients of topic ( t )</td>
</tr>
<tr>
<td>( w_{st} )</td>
<td>Model coefficients of topic ( t ) for source ( s )</td>
</tr>
<tr>
<td>( \alpha, \beta, \gamma )</td>
<td>Regularization parameters</td>
</tr>
</tbody>
</table>

### 5.2.2. Model Definition

This sub-section details how we utilise multiple types of user data on Twitter to better model the problem of user expertise inference. In the last chapter, we proposed a Sentiment-weighted and Topic Relation-regularized Learning (SeTRL) model to address this problem. SeTRL first builds the feature vector of a user based on the user’s tweets and utilises the sentiment intensity contained in the tweets to weight the features of each user. Then by using linear regression, a base model is built to jointly learn the expertise of users on multiple topics. Meanwhile, SeTRL exploits the relatedness between expertise topics to optimize inference, which is characterized by an undirected graph \( G \) with \( E \) edges. It encodes this relatedness information in the base model through model regularization. The SeTRL model is represented by Eq.(4.5) in Section 4.2.3.

However, SeTRL is heavily dependent on abundant content posted by the user in order to perform expertise inference. This model will struggle when a user has not posted a sufficient number of tweets. As discussed in Section 5.1, in practice, a large proportion of Twitter users do not actively post tweets, or have never posted any tweets. In these cases, we need to seek other information to help infer a user’s expertise. In addition to posting tweets, a Twitter user can also interact with other users through various other activities. These activities may also provide effective evidence about the user’s expertise information. As explained in Section 5.1, authors in [We10] exploited the “following” relationships between users to identify topically influential Twitter users and they believe that a user tends to follow other users with similar topical interests; the authors in [Gh12]
observed that list data on Twitter provides valuable semantic cues to the topics of expertise of the users on the list, so they proposed exploiting this data to mine topic expertise on Twitter. Additionally, different users on Twitter exhibit different behaviours and habits. Some users like posting tweets and there are also users who mainly just follow others and read content that interests them. Therefore, it is important to jointly exploit multiple types of user data for expertise inference. This will benefit cases where only certain types of user data are available. In this research, we propose incorporating multiple types of data associated with the user into the process of user expertise inference through the loss function:

$$ \sum_{i=1}^{N} \sum_{t=1}^{T} \frac{1}{2} (y_{ti} - \sum_{s=1}^{S} \frac{1}{|S|} x_{si} w_{st})^2 $$

(5.1)

where $x_{si}$ is the feature vector of user $i$ defined from the data that is associated with the user through the $s^{th}$ relationship; $w_{st}$ is the model parameters of the $s^{th}$ data source part of expertise topic $t$; $|S|$ is the total number of relationship types considered in the model.

Thus, a base model has been built that can infer a user’s expertise by combining multiple types of data associated with the user. This model does not distinguish between the different types of user data in the process of expertise inference. Whereas in reality, out of the multiple types of data associated with a user, some have more related information than others to the topics we are trying to infer. So, the model may only be able to identify discriminative features from certain types of user data. In this case, using different types of user data to infer a user’s expertise on a topic may have inconsistent results. In fact, as discussed in Section 5.1, most Twitter users do not have sufficient information for all the types of data (i.e. tweets, followers, friends and lists). This could result in a low prediction accuracy in the base model. In this work, we use the following regularizer to penalize the disagreement between different sources:

$$ \sum_{t=1}^{T} \sum_{s_1=1}^{S} \sum_{s_2=1}^{S} \sum_{s_1 \neq s_2}^{S} \left| x_{s_1i} w_{s_1t} - x_{s_2i} w_{s_2t} \right|^2 $$

(5.2)

where $s_1$ and $s_2$ are any two different types of user data from $S$. This regularization term tries to model the inference consistency among different types of data associated with the user. It aims to ensure the model has good prediction accuracy regardless of the user having, or not having, all four different types of data. Here, we consider a specific example to illustrate the usefulness of this regularization process. In the training data,
there are Twitter users who have expertise in the area of “deep learning” and posted many
tweets about this topic such as:

“Theano is a good deep learning framework for researchers”.

At the same time, some users who have expertise in this topic are also included in a few
lists with the list names like “BigData expert” and “Machine learning”. Based on this
fact, the model in Eq.(5.1) would identify features like “deep learning”, “theano” in the
user’s tweets as the most discriminative features for the expertise topic “deep learning”
and assign high weights to them. In comparison, features like “bigdata” and “machine
learning” in the user’s list data would receive a lower weight but they are actually
important features for this topic too. It can be expected that the learned model for the
expertise topic “deep learning” would perform well if there are sufficient user posted
tweets available, while it may produce unsatisfying prediction results if we only have the
user’s list data. The regularizer in Eq.(5.2) tries to balance the weights on the important
features from different types of user data. In this example, it will decrease the weights for
features like “deep learning” and “theano” in the user’s tweets and increase the weights
for features like “bigdata” and “machine learning” in the user’s list data by looking at the
prediction accuracy of the model using only the tweet or list data of the training data
users. Thus, the learned model ensures it will deliver good prediction performance both
in cases where the user has one of the two types of data, and those where they have both.

Now, we can construct our multi-Data and Topic relatedness Combined (D^TCom)
learning model based on the above formulation. By substituting Eq. (5.1) and Eq. (5.2)
into the SeTRL model Eq. (4.5), we have the following optimization problem:

\[
\min_W \sum_{t=1}^{N} \sum_{t=1}^{T} \frac{1}{2} \left( y_{ti} - \sum_{s=1}^{S} \frac{1}{|S|} x_{si} w_{st} \right)^2 \\
+ \gamma \sum_{l=1}^{N} \sum_{t=1}^{T} \sum_{s_1=1}^{S} \sum_{s_2 \neq s_1}^{S} |x_{s_1 t} w_{s_1 t} - x_{s_2 t} w_{s_2 t}|^2 \\
+ \alpha \sum_{e=1}^{E} |w_{e(1)} - w_{e(2)}|^2 + \beta |W|_1
\]  

(5.3)

where \(\gamma\) is the regularization parameter to control the contribution of the inference
consistency of different types of user data.
Similar to the last chapter, this chapter of work also adopts the Accelerated Proximal Gradient method [Ji09] to learn the proposed model, i.e. to estimate a parameter configuration of \( W \) that minimizes the objective function Eq.(5.3). It uses a linear combination of the previous two points as the search point to achieve high convergence speed.

### 5.3. Dataset Construction and Experiment Preparation

To evaluate the proposed D⁰TCom learning model, we need a dataset that contains the expertise information of a large number of Twitter users. The last chapter described the construction of such a dataset using the popular question and answering service Quora. This is based on the fact that many Quora users explicitly provide both their expertise topics and Twitter account information in their profiles. So, the same person’s Twitter account and expertise topics were harvested, and these expertise topics were used as the ground truth of the expertise of the Twitter user in the experiments. The dataset contains 10,856 Twitter users and 149 expertise topics from which each user has, at least, knowledge of one topic. We reuse this dataset in the experiments detailed in this chapter.

While in the last chapter, the experiments only used the user’s tweets for expertise inference, this part of the PhD research needs to use other types of data of the user, i.e. friends, followers and list data. So, we harvested this additional data of each user in the dataset, if it was available, using the official Twitter API. In terms of both the follower and friend data, the Twitter API imposes a limit of 5000 users, which are collected for each user.

Figures 5-1, 5-2 and 5-3 show the distribution of the number of collected friends, followers and lists of all the users in the dataset respectively. For example, as shown in Figure 5-1, we failed to collect any friend data for 904 users (8.3%) from the 10,856 Twitter users. This is as a result of one of the following two scenarios: (i) the user does not follow any other users; (ii) or the user does not make their “following” data publicly available. There are 1,539 users (14.2 %) who have over 1500 friends and most of the users (5,905 users, 54.4%) have between 1 and 500 friends in the dataset. In comparison, list membership is much less frequent for the average Twitter user, as shown in Figure 5-3. In the dataset there are 1,733 users (15%) who are not a member of any list, and over half of the users are included in less than 20 lists.
Figure 5-1: Distribution of the friend numbers of Twitter users in the dataset

Figure 5-2: Distribution of the follower numbers of Twitter users in the dataset
In this experiment, the profile information of the friends and followers of a Twitter user is used as an input for the inference task. A user profile on Twitter includes attributes such as user name, location and a short bio. There are two main reasons why we did not include the tweets posted by the friends and followers for inference. First, there are millions of friend and follower users in the dataset. The Twitter API limitation prevents us from harvesting the tweets of that many users within a reasonable time-frame. Second, a user’s own tweets are already noisy with regard to reflecting on his/her expertise [Wa12]. Therefore, including the tweets of their friends and followers is likely to introduce even more noise into the process. In comparison, a user’s profile provides direct personal information about him/her. Although, at an individual level, the profile may not contain any expertise information, or may even have missing elements, aggregating the profiles of a group of users could capture the main characteristics of the group [He13]. Thus, the text information in the profiles of all friends or followers of a user are combined as an input document (called friend document or follower document) for inferring the user’s expertise. In terms of the list data, the name and description information of the lists of the user are combined as the input document (list document). Correspondingly, the combination of the user’s posted tweets is called the tweet document of the user. In the experiments, we combine multiple documents of a user, such as tweet document and friend document, as one input document for testing the effectiveness of the combination of multiple types of user data in user expertise inference.

![Figure 5-3: Distribution of the list numbers of Twitter users in the dataset](image-url)
This work considers two methods for constructing the user features with the input document. 1) **Unigram features**: This method uses the bag-of-words features of the document as the user feature space. 2) **Latent topic features**: LDA is applied to generate the latent topic distributions of the input documents of users, which are used as the user features. In the construction of experiment samples for each expertise topic, we only take the users with documents of over 100 terms as valid samples. Then, balanced positive and negative samples of each topic are randomly chosen from the dataset, two thirds of which are used as the training set. The remaining one third is the test set for the topic. In the experiment, we do not consider expertise topics with less than 50 samples. As there are a large number of users in the dataset who do not have any, or have limited, list data, 64 expertise topics meet the experiment requirements when testing with the user list data alone, while experiments with tweet, friend and follower data respectively cover all the 149 topics.

### 5.4. Experiments and Results

This section first outlines the metrics used to measure the performance of the various methods in the experiments. It then analyses, in detail, the experimental results of the methods which use a single type of user data, and those which combine multiple types of user data.

#### 5.4.1. Evaluation Metrics

Like the previous two chapters, the four metrics: accuracy, precision, recall and F1-score are used to measure the performance of methods in this work. Specifically, we use the average score of each of the four measurements on all the tested topics to examine the performance of various inference methods. In the experiments, a standard 5-fold cross validation on the training data is performed, as in the previous chapter, to select the regularization parameters $\alpha$, $\beta$ and $\gamma$.

#### 5.4.2. Performance of Different Types of User Data

The problem of user expertise inference has been studied using the user’s tweets in the last chapter. It examined the performance of the Support Vector Machine (SVM) method with unigram features of user tweet documents. It experimented on various schemes of weighting the user unigram features and discovered that the tweet sentiment-based weighting scheme performed best. Using this weighting scheme, it then takes into
consideration the topic relatedness for user expertise inference, which is the proposed SeTRL model in the last chapter. In this chapter of work, a similar method is used to examine the performance of each of the other three types of user data (i.e. friends, followers and lists). Specifically, we first use SVM to examine the performance of each type of user data with different user features (i.e. unigram features and latent topic features) and different weighting schemes. The aim is to identify an appropriate user feature space and the optimal feature weighting scheme. Then, we incorporate topic relatedness into the inference model (i.e. model Eq. (4.5) but with the identified method of constructing user feature vectors) to further examine the performance of each type of user data. Note that the topic co-occurrence based method used in the last chapter is applied to construct the topic relation graph in the experiments. Below, we will present and analyse the experimental results using two methods of user feature construction: unigram features and latent topic features.

**Unigram Features**: The following schemes are used to weight unigram features of a user document in the experiments:

1. **BI**: This scheme uses a binary value to represent the feature value. It means that the feature value is set to 1 if this user document has this feature (term), otherwise it is set to 0;

2. **TF**: This scheme uses the frequency of the feature terms occurring in the user document as its weight;

3. **TFIDF**: This scheme uses the Term Frequency – Inverse Document Frequency (TF-IDF) algorithm [Ai03] to assign weights to the user features, where all the user documents are taken as a document set.

For each type of user data, we use the above three schemes to build user feature vectors, and employ SVM for user expertise inference. Table 5-2 presents the performance of the SVM method with the input of each type of user data and each weighting scheme. The SVM-Sen method adopts the tweet sentiment based weighting scheme proposed in the last chapter for user expertise inference using tweets (the best weighting scheme identified). The experimental results first show that the TF weighting scheme outperforms the other two schemes on all the three types of user data (apart from the recall using followers). For example, when using friend data for user expertise inference, the SVM-TF achieves about 5% (F1-score) improvement, when compared to SVM-TFIDF. This
demonstrates that commonly used terms in all the documents are also useful for user expertise inference on Twitter and decreasing their importance will harm the prediction accuracy. So, the TF weighting scheme is applied to the experiments below that consider the topic relatedness or the use of a combination of multiple types of user data for expertise inference.

Table 5-2: Performance of SVM method with each type of user data for user expertise inference (%)

<table>
<thead>
<tr>
<th>Data Type</th>
<th>Method</th>
<th>Recall</th>
<th>Precision</th>
<th>Accuracy</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tweets</td>
<td>SVM-Sen</td>
<td>72.10</td>
<td>70.88</td>
<td>69.75</td>
<td>69.76</td>
</tr>
<tr>
<td>Friends</td>
<td>SVM-BI</td>
<td>67.84</td>
<td>67.11</td>
<td>65.46</td>
<td>64.75</td>
</tr>
<tr>
<td></td>
<td>SVM-TF</td>
<td>68.37</td>
<td>75.43</td>
<td>72.23</td>
<td>70.77</td>
</tr>
<tr>
<td></td>
<td>SVM-TFIDF</td>
<td>66.09</td>
<td>71.89</td>
<td>67.18</td>
<td>65.53</td>
</tr>
<tr>
<td>Followers</td>
<td>SVM-BI</td>
<td>68.55</td>
<td>61.54</td>
<td>62.67</td>
<td>62.72</td>
</tr>
<tr>
<td></td>
<td>SVM-TF</td>
<td>61.64</td>
<td>74.23</td>
<td>69.37</td>
<td>66.08</td>
</tr>
<tr>
<td></td>
<td>SVM-TFIDF</td>
<td>57.16</td>
<td>71.16</td>
<td>61.73</td>
<td>56.45</td>
</tr>
<tr>
<td>Lists</td>
<td>SVM-BI</td>
<td>64.61</td>
<td>69.76</td>
<td>66.89</td>
<td>64.46</td>
</tr>
<tr>
<td></td>
<td>SVM-TF</td>
<td>66.04</td>
<td>74.47</td>
<td>70.18</td>
<td>68.29</td>
</tr>
<tr>
<td></td>
<td>SVM-TFIDF</td>
<td>60.13</td>
<td>77.72</td>
<td>69.79</td>
<td>64.98</td>
</tr>
</tbody>
</table>

In addition, it is also observed from Table 5-2 that using friend data or list data can achieve better performance than using follower data for user expertise inference. For example, SVM-TF achieves a 70.77% F1-score with friend data, while it gets a 66.08% F1-score with follower data. This difference indicates that information from a Twitter user’s friends or lists can more effectively reflect the user’s expertise. This confirms the intuition that a user tends to follow other users to reach the content she/he is interested in, and that a user is usually included in lists which contain some information about her/him. However, a user typically has no control over his/her followers and anyone can follow a user without his/her permission. This could introduce significant noise, such as spam users or advertisement users, to the expertise inference process. Furthermore, we use the TF weighting scheme to build user feature vectors and consider the topic relation information for user expertise inference, which is the topic relation-regularized learning (TRL) model from the last chapter. As shown in Table 5-3, this validates again that friend and list data are more effective in inferring a Twitter user’s expertise information.
Table 5-3: Performance of TRL method with each type of user data and unigram features (%)

<table>
<thead>
<tr>
<th>Data Type</th>
<th>Method</th>
<th>Recall</th>
<th>Precision</th>
<th>Accuracy</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tweets</td>
<td>SeTRL</td>
<td>82.49</td>
<td>78.76</td>
<td>79.65</td>
<td>80.08</td>
</tr>
<tr>
<td>Friends</td>
<td>TRL-TF</td>
<td>80.99</td>
<td>80.24</td>
<td>79.82</td>
<td>80.09</td>
</tr>
<tr>
<td>Followers</td>
<td>TRL-TF</td>
<td>76.33</td>
<td>78.92</td>
<td>77.51</td>
<td>77.32</td>
</tr>
<tr>
<td>Lists</td>
<td>TRL-TF</td>
<td>78.21</td>
<td>79.95</td>
<td>78.90</td>
<td>78.89</td>
</tr>
</tbody>
</table>

Latent Topic Features (LTF): The LDA algorithm generates a probability distribution over the latent topics to represent the topic distribution of a user document. The probability score can be naturally used as the weight of a latent topic feature of a user. Although there is no need to select an appropriate feature weighting scheme, we still need to identify an optimal latent topic dimension. Therefore, we observe the performance of SVM with different topic dimensions for the optimal dimension selection. Figures 5-4, 5-5, 5-6 and 5-7 show the accuracy and F1 performance of SVM with the varying dimensions of latent topics for each type of user data respectively. The results illustrate that its performance rises as the topic dimension increases and reaches the best performance when the dimension is set at approx. 100. It then starts to decline for each type of user data. This means that about 100 latent topics can best represent the content richness of the input documents for our expertise inference task. Less or more topics could result in an underrepresentation or overrepresentation of the input content respectively. Thus, we select 100 as the optimal topic dimension and use it for the next experiments. From the experimental results, it can also be observed that the SVM with latent topic features achieves a similar or slightly better performance for each type of user data (considering the optimal topic dimension), compared to the SVM with unigram features. For example, when using follower data for user expertise inference, SVM-TF achieves a 66.08% of F1 score, while the SVM with latent topic features gets a slightly higher 67.3% of F1 score (topic dimension is set as 100). This improvement could benefit from the representation of a general topic distribution of the user. It may help to alleviate the impact of noise in hundreds of thousands of unigram features but it also loses much important information about the user as you will see from the experimental analysis below.
Figure 5-4: Performance of SVM with tweet data and latent topic features

Figure 5-5: Performance of SVM with friend data and latent topic features
Similar to the unigram features, we also examine the performance of TRL with the user latent topic features. As shown in Table 5-4, the inclusion of the topic relation information does not remarkably improve its performance, when compared to the SVM method. This
result indicates that the latent topic distribution of a user can effectively reflect the knowledge information of a user on a specific expertise topic. But this feature representation may lose some detailed information about the user from the original user document, which could limit its potential for exploitation to further improve user expertise inference. Specifically, the relation information is most useful in the case where there is only limited evidence available to infer a user’s expertise on a certain topic. However, this limited information about the user usually cannot be reflected from the latent topic distribution.

<table>
<thead>
<tr>
<th>Data Type</th>
<th>Method</th>
<th>Recall</th>
<th>Precision</th>
<th>Accuracy</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tweets</td>
<td>TRL-LTF</td>
<td>74.56</td>
<td>70.30</td>
<td>71.51</td>
<td>72.06</td>
</tr>
<tr>
<td>Friends</td>
<td>TRL-LTF</td>
<td>73.95</td>
<td><strong>71.16</strong></td>
<td><strong>71.89</strong></td>
<td><strong>72.22</strong></td>
</tr>
<tr>
<td>Followers</td>
<td>TRL-LTF</td>
<td>72.31</td>
<td>69.98</td>
<td>70.50</td>
<td>70.87</td>
</tr>
<tr>
<td>Lists</td>
<td>TRL-LTF</td>
<td>73.90</td>
<td>70.35</td>
<td>71.27</td>
<td>71.53</td>
</tr>
</tbody>
</table>

Overall, the above experimental results and analysis demonstrate that each type of user data is useful for user expertise inference but with varying performance. In terms of the construction of user features, although both schemes can deliver good inference performance, the unigram features keep more information about the user from the original data and leave space for advanced methods to further improve the performance. So, we will use the unigram features for the experiments in the next subsection that describes the use of the combination of multiple types of user data for expertise inference.

### 5.4.3. Performance of Combinations of Different Types of User Data

This sub-section examines the performance of different methods using the combination of different types of user data for expertise inference. Table 5-5 compares the performance of our D^TCom learning model with that of SVM and SeTRL using all four types of user data. Note that for SVM and SeTRL, the features generated from different data sources are directly concatenated as a single feature vector. Experimental results show that SVM achieves the worst performance, which is even lower than that of SVM using one type of user data alone. This could be due to over-fitting, as much inconsistent information is considered in the learning process. It is also shown that D^TCom
significantly outperforms SeTRL. It verifies the significance of taking source consistency into consideration when using multiple types of user data for expertise inference.

Table 5-5: Performance of different methods with the combination of the four types of user data (%)

<table>
<thead>
<tr>
<th>Methods</th>
<th>Recall</th>
<th>Precision</th>
<th>Accuracy</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM</td>
<td>57.26</td>
<td>73.87</td>
<td>67.67</td>
<td>62.64</td>
</tr>
<tr>
<td>SeTRL</td>
<td>83.93</td>
<td>80.56</td>
<td>81.01</td>
<td>82.13</td>
</tr>
<tr>
<td>DTCom</td>
<td>91.75</td>
<td>83.35</td>
<td>85.72</td>
<td>86.80</td>
</tr>
</tbody>
</table>

Table 5-6: Performance of DTCom with various combinations of different types of user data (%)

<table>
<thead>
<tr>
<th>Data Type Combinations</th>
<th>Recall</th>
<th>Precision</th>
<th>Accuracy</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tweets+Friends</td>
<td>87.23</td>
<td>82.65</td>
<td>84.24</td>
<td>85.16</td>
</tr>
<tr>
<td>Tweets+Followers</td>
<td>84.62</td>
<td>80.71</td>
<td>81.76</td>
<td>82.67</td>
</tr>
<tr>
<td>Tweets+Lists</td>
<td>85.93</td>
<td>81.23</td>
<td>83.10</td>
<td>83.93</td>
</tr>
<tr>
<td>Tweets+Friends+Followers</td>
<td>88.09</td>
<td>81.87</td>
<td>84.76</td>
<td>84.90</td>
</tr>
<tr>
<td>Tweets+Friends+Lists</td>
<td>90.73</td>
<td>83.28</td>
<td>85.11</td>
<td>86.32</td>
</tr>
<tr>
<td>Tweets+Followers+Lists</td>
<td>86.58</td>
<td>82.47</td>
<td>84.03</td>
<td>84.15</td>
</tr>
<tr>
<td>Tweets+Friends+Followers+Lists</td>
<td>91.75</td>
<td>83.35</td>
<td>85.72</td>
<td>86.80</td>
</tr>
</tbody>
</table>

Furthermore, we also conduct experiments to observe the performance of our model with various combinations of the four types of user data, as shown in Table 5-6. The results show that aggregating all four types of user data using our model for expertise inference always achieves better performance than using any single type of user data, and the more data about the user we incorporate, the better performance it can achieve. For instance, our model achieved the best F1 score of 86.8% in the experiment that exploits all four types of user data. It is also noted that the combination of the two types of user data, tweets and friends, can obtain satisfying performance, while the combination of the three types of user data: tweets, followers and lists achieves lower performance. This could be due to the fact we discussed previously that tweet and friend data are more effective in reflecting a user’s expertise, while follower and list data are either more noisy or sparse for many users.
5.5. Chapter Summary

Based on the work of the last chapter, this chapter further studied the problem of user expertise inference on Twitter, with a focus on the exploitation of various types of data associated with the user. A multi-Data and Topic relatedness Combined (D^TCom) learning model was proposed that could infer the user’s topical expertise under the influence of multiple types of user data. The D^TCom considers the inference consistency of different types of user data in the process of learning. It aims to optimize the model learning based on the assumption that a user’s expertise reflected by each type of data associated with him/her should be similar. Experiments are conducted on the same Twitter dataset whose construction was outlined in the last chapter, but with more activity data harvested for each user. In the experiments, four different types of user-related data are tested: tweets, friends, followers and lists. Detailed experimental analysis demonstrates that each type of user data is effective for user expertise inference, with variation in performance. Experimental results show that our D^TCom learning model can better make use of the various data sources associated with the user, when compared with several baseline approaches, and that combining the data sources in a single model produces the best expertise inference performance.

The experimental work described in this chapter was published in the following papers:


6. A Case Study: Finding Right Answerers in CQA Sites

6.1. Introduction

The previous three chapters presented different modelling approaches to infer a SNS user’s expertise information from their social media content, i.e. language expertise inference from static social profiles (Chapter 3), and topical expertise inference from social activities on Twitter (Chapter 4 and Chapter 5). In order to study the benefit of the proposed expertise inference and modelling approaches for real-world applications, this chapter will describe the utilisation of the inferred expertise of SNS users in a Community Question Answering (CQA) service.

The CQA services allow their users to answer questions asked by other users. CQA sites such as Quora and Yahoo! Answers have attracted millions of users over the last couple of years. According to statistics in April 2017\(^\text{26}\), Quora has over 190 million unique visitors a month with over 40 million users from the US alone. To offer a better CQA service, one of the most important aspects is to ensure that newly asked questions can be efficiently solved. To address this problem, finding users who can answer the question, i.e. who have knowledge of the topics contained in the question, is key. In this chapter of the PhD thesis, we will attempt to exploit the expertise information inferred from users’ social media content to help Answerer Finding (AF) on the CQA site, Quora.

As introduced in previous chapters, many Quora users provide their social media accounts on their profiles, e.g. Twitter and LinkedIn. This enables us to access the Quora user’s public LinkedIn profiles and their social activities on Twitter. Thus, their expertise information can be inferred correspondingly from our proposed inference approaches. However, this work focuses on the exploitation of the topical expertise inferred from the user’s Twitter content, not including the language information inferred from their LinkedIn profiles. This is primarily because there is not a significant volume of multilingual question content on Quora at this time, which prevents us from constructing the required experimental settings, and to the best of our knowledge, there are no appropriate public datasets available that suit our experimental requirements. Thus, the utilisation of inferred language information in real-world application scenarios, e.g.

\(^{26}\) http://expandedramblings.com/index.php/quora-statistics/
multilingual question answering or personalised information retrieval, will form part of planned future work.

As discussed in detail in Section 2.3, previous studies primarily rely on the user’s rich answering history to model the user’s expertise information, which are then used to match with new questions. In practice, a large proportion of users in CQA sites have never answered any questions or only answered a small number [Sr15]. These users are referred to as cold-start users. For example, statistics from the harvested Quora dataset in this research show that over half of the users answered less than 10 questions (see Section 6.3). This means that User Answering History (UAH) based approaches immediately excluded most of the users in the CQA site when selecting question answerers. However, as introduced in Chapter 4, Quora allows users to specify their expertise topics in their profiles in order to more effectively match questions with their potential answerers. This research, as detailed in previous chapters, proposed to infer their topical expertise by using the user’s social media content. Intuitively, the inferred topical expertise of users could be also exploited to represent their knowledge areas and distinguish the user’s ability to answer a question from others. If this intuition is shown to be true, then users who have linked to their SNS accounts can still be included in AF, even if they did not previously answer many questions or specifically provide their expertise topics. This will help CQA services to expand the search scope in AF and improve the chance of finding appropriate answerers for new questions. Therefore, to allow a CQA service to benefit from this user information, this chapter describes research which aims to answer the following specific question:

*Can the inferred user expertise from users’ social media content be used to facilitate the answerer finding service on a CQA site?*

Specifically, the contributions of this chapter of work are twofold:

1. Based on a traditional UAH based AF approach, this work proposes an approach that weights a user’s ability to answer a question by utilising the user’s expertise topics. It should be noted that for comparison purposes, the user’s specified expertise topics (ground truth) are also applied to AF in the study.

2. This work creates a dataset based on the Quora data used in the previous chapters to simulate the application scenario of AF on Quora. Experiments conducted on the dataset demonstrate that the user’s expertise topics (both inferred and specified) are
a useful information source for AF, and the User Expertise Topic (UET) based approach can achieve prediction accuracy as good as that of the UAH based approach.

The rest of this chapter is organized as follows. Section 6.2 describes in detail two answerer finding approaches which use two different information sources: user answering history and user expertise topics; Section 6.3 introduces the construction of the experimentation dataset; Section 6.4 details the experimental evaluation method; Section 6.5 gives the experimental results and analysis; Section 6.6 concludes this chapter.

6.2. Answerer Finding on CQA Sites

This section introduces two answerer finding approaches which are based on two different information sources of the user: questions answered by the user and expertise topics of the user (specified or inferred from his/her social media content).

6.2.1. User Answering History Based Approach

As reviewed in Section 2.3, previous studies primarily rely on a user’s past answering activities to build potential answerers’ expertise models, and estimate the matching score between a question and the user’s expertise model for AF in CQA sites. Thus, potential answerers can be ranked according to the estimated matching score for a new question. In the literature, a variety of approaches have been proposed to address the problem. They studied numerous user expertise modelling approaches, such as topic model-based approaches [Gu08], language model-based approaches [Li10a], and their corresponding matching strategies. However, this research is not aiming to build a superior AF matching strategy based on the proposed topical expertise modelling approach. Instead, we adopt a conventional language model-based approach as the AF baseline, for two main reasons:

(1) The purpose of this piece of research is to investigate the usefulness of the information in the inferred user expertise topics in the context of answerer finding. It is not intended for this research to propose a new matching strategy which utilises this information, when compared to existing approaches. We leave this aspect of the research as the focus of future work.

(2) This work focuses on the cold-start scenario, i.e. users who have authored few or no previous answers, where existing approaches will struggle to model the user’s expertise. It aims to demonstrate that the user’s inferred expertise topics are an
effective information source and can be used to find appropriate answerers for new questions.

Specifically, the adopted language model-based approach is built on the TF-IDF weighting algorithm [Li10b, Ti13]. The idea of this approach is that, given a new question, users who have answered similar questions in the past should have the ability to answer this question. In this approach, all the questions answered by a user are combined (both question and answer text bodies included) as a document, which is the user’s expertise information source. By applying the TF-IDF weighting algorithm on the vector space model, each document is represented as a vector of weighted features which is the user expertise model. In each vector, the features are the words which appear in the user document and are weighted by TF-IDF. Given a user expertise model $\theta_u$ and a new question $q$, the approach ranks the users based on the cosine similarity between $\theta_u$ and $q$. The detailed formula is as follows:

$$s(\theta_u, q) = \frac{\sum_w TFIDF(w, \theta_u) \times TFIDF(w, q)}{\sqrt{\sum_w TFIDF(w, \theta_u)^2} \times \sqrt{\sum_w TFIDF(w, q)^2}}$$

(6.1)

where $w$ denotes the words that appeared in both user $u$’s expertise model and question $q$; $TFIDF(w, \theta_u)$ is the TF-IDF weight of word $w$ in $\theta_u$; $TFIDF(w, q)$ is the TF-IDF weight of word $w$ in $q$.

### 6.2.2. User Expertise Topic Based Approach

In a cold start context the user answering history based approach does not have sufficient answering activities to build a user expertise model. In this case, this work proposes to utilise the user’s inferred expertise topics for user expertise modelling. As stated previously, the aim of this research is to demonstrate the usefulness of user expertise topics in finding question answerers, rather than proposing a new user-question matching strategy. Therefore, to make it directly comparable to the baseline approach and simplify the evaluation process, we design the answerer finding approach using expertise topics on the basis of the baseline approach. The detailed steps are as follows:

1. There are previously answered questions in a CQA site that have been categorized by specific knowledge topics. The topics of these questions share the same folksonomy as the user’s expertise topics. These questions are gathered to form a
question pool. Details on the construction of this question pool are introduced in the following section.

(2) Given a cold start user, we assume the user’s expertise topics are provided in the profile or can be inferred from her/his social media content. For each topic associated with the user, a certain number of questions that have been categorised as being in that topic are selected from the question pool (constructed in step 1). The selected questions are used to simulate the cold start user’s answering history.

(3) Similar to the baseline approach, the simulated answering questions of the cold start user are then combined as a document and used to build the user’s expertise model for answerer finding.

This approach relies on the expertise topics of the user for AF without the use of the user’s actual answering activities. If this approach can effectively find appropriate answerers for new questions from cold start users, it validates that the user’s expertise topics inferred from their social activities are a useful information source for user expertise modelling and can help to include more users in the process of AF. More detail on this approach and the experiments will be introduced in Section 6.4.

6.3. Dataset Construction

The experiments described in this chapter are considered a continuation of the experiments of Chapter 5. They will continue to be conducted on the Quora dataset used in both Chapter 4 and Chapter 5, but with the research target of trying to identify more potential answerers for new questions. Therefore, we extended the dataset to 24,285 Quora users, which includes the previous 10,856 users, in order to expand the user search pool. All these users have provided both their Twitter account information and expertise topics on their Quora profiles, which have been harvested. Note that only the 10,856 users used in the last two chapters have expertise of the 149 topics selected. To build the testing cases, we also needed to collect the questions these users have answered. So, we harvested each user’s answer page on Quora, which lists all the questions the user has previously answered. Figure 6-1 gives an example of a Quora user’s answer page. It should be noted that if the answer provided by the user to a question is too long, the answer body will not be completely displayed in the user’s answer page, such as the second question in Figure 6-1. In this case, we extracted the URL link to the full answer from the page and harvested the full answer body correspondingly. Figure 6-2 gives an example of a page that displays
the full answer to a long question. In terms of those questions fully displayed in the user’s answer page, we directly parsed the HTML page and extracted both their question title and answer body. A maximum of 100 questions have been collected for each user in the experiments. Please note that those users for who 100 questions were collected in the dataset usually answered over 100 questions on Quora.

Figure 6-1: An example of a Quora user’s answer page

Figure 6-2: An example of the page of the full answer to a question
In total, 591,654 questions answered by the 24,285 users have been collected, among which 440,994 were harvested from external answer pages. Table 6-1 summarises the distribution of the number of questions collected per user. It indicates that about 43% of the users answered less than 5 questions and only about 10% of the users have answered over 100 questions. Thus, in practice, a big proportion of users on Quora can be considered cold start users, which are ignored by existing user answering history based approaches. This demonstrates the significance of exploiting other information sources related to the user for AF, e.g. topical expertise inferred from the user’s social media content.

Table 6-1: Distribution of the number of questions collected per user

<table>
<thead>
<tr>
<th>Questions Collected</th>
<th>Questions Collected</th>
</tr>
</thead>
<tbody>
<tr>
<td>≥ 1</td>
<td>19,418</td>
</tr>
<tr>
<td>≥ 5</td>
<td>13,939</td>
</tr>
<tr>
<td>≥ 10</td>
<td>11,398</td>
</tr>
<tr>
<td>≥ 20</td>
<td>8,083</td>
</tr>
<tr>
<td>≥ 30</td>
<td>6,195</td>
</tr>
<tr>
<td>≥ 40</td>
<td>5,061</td>
</tr>
<tr>
<td>≥ 50</td>
<td>4,296</td>
</tr>
<tr>
<td>≥ 60</td>
<td>3,736</td>
</tr>
<tr>
<td>≥ 70</td>
<td>3,265</td>
</tr>
<tr>
<td>≥ 80</td>
<td>2,898</td>
</tr>
<tr>
<td>≥ 90</td>
<td>2,611</td>
</tr>
<tr>
<td>= 100</td>
<td>2,375</td>
</tr>
</tbody>
</table>

As mentioned in the previous section, it was necessary to construct a question pool for the proposed UET based approach. All the questions collected in the pool should have their knowledge categorization information provided. Fortunately, each Quora question has a corresponding page on which all the answers given by different users are listed and users can manually edit which knowledge topics the current question falls into. Figure 6-3 gives an example of a Quora question page with its knowledge topics annotated. Therefore, to construct the desired question pool, we only need to go to the question page of the selected questions and harvest their knowledge topics if available. From those users in the dataset who answered over 100 questions on Quora, their last 15 questions with knowledge topic information available were selected to create the question pool, which
are 26,276 questions in total. There are two main reasons why we only used the last 15 questions of the users who answered over 100 questions:

Figure 6-3: An example of a question page with knowledge topics

(1) In the experiments, the questions of those users who answered over 100 questions will be used to evaluate the performance of the AF approaches and the topic information of the questions collected here will also be used to create the testing cases (see details in the next section);

(2) The selected questions are enough to cover the testing of expertise topics in the experiments. Harvesting question pages for all the questions would consume a significant amount of time.

In summary, the dataset used in this research comprises 24,285 Quora users with both their Twitter accounts and expertise topics provided. In addition, it also includes 591,654
questions answered by these users and from among these, the knowledge topics of 26,276 questions are also harvested to form a question pool.

6.4. Evaluation Method

6.4.1. Ground Truth Answerers Construction

To evaluate the effectiveness of the answerer finding approach using expertise topics, the key is to demonstrate it can at least achieve a prediction accuracy as good as that of the UAH based approach. In other words, given a new question, experimental results are expected to show that the quality of answerers selected by the UET based approach should be as good as that of answerers selected by the UAH based approach.

Similar to previous studies [Li10b, Gu08], experimental evaluation in this work is based on the assumption that users who actually answered a question are considered as the target answerers to that question. Therefore, questions from the constructed dataset are selected to build the testing set and the users who actually answered the question in the dataset are taken as the ground truth answerers. Also, to better evaluate the UET based approach, the testing questions are selected by the group of expertise topics. This can help to reduce the effect of noise on experimental results and different answerer finding approaches can be evaluated on different topics of questions, which will be explained in detail later in the next sub-section. Specifically, testing questions for each expertise topic are selected by the following steps:

1. An expertise topic is selected from the 149 topics in the dataset;

2. From the last 15 questions of the users with over 100 questions collected (i.e. the 2,375 users in the dataset), those questions categorised in the expertise topic are selected as the testing questions for that topic. For each selected question, the user who answered the question is used as the ground truth answerer;

3. All the selected question-user pairs are used as the testing samples for the topic.

In the selection of testing questions, we only include the users who answered over 100 questions. This is because we first need to validate the effectiveness of the user answering history based approach and observe to what extent it can find appropriate answerers for given questions. Once this is complete, a comparison can be performed with the prediction accuracy of the expertise topic based approach (details of this evaluation are given in the next sub-section). As a result, it is necessary to have sufficient answering
activities for the selected ground truth answerers. As explained previously, one of the reasons why the testing samples are only selected from the user’s last 15 questions is due to the harvesting time limitation. Additionally, it aims to best simulate the real application scenario, i.e. finding answerers for new questions based on the user’s past answering activities. At the same time, we also hope to include as many as possible different users in the testing samples for each topic.

Table 6-2: Statistics of the testing samples by expertise topics

<table>
<thead>
<tr>
<th>Topic Name</th>
<th>No. of Users</th>
<th>No. of Question-User Pairs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sales</td>
<td>39</td>
<td>16</td>
</tr>
<tr>
<td>Machine Learning</td>
<td>49</td>
<td>30</td>
</tr>
<tr>
<td>Video Games</td>
<td>33</td>
<td>14</td>
</tr>
<tr>
<td>Music</td>
<td>110</td>
<td>45</td>
</tr>
<tr>
<td>Recruiting</td>
<td>26</td>
<td>14</td>
</tr>
<tr>
<td>Photograph</td>
<td>67</td>
<td>15</td>
</tr>
<tr>
<td>Philosophy</td>
<td>101</td>
<td>13</td>
</tr>
<tr>
<td>Islam</td>
<td>20</td>
<td>15</td>
</tr>
<tr>
<td>Startups</td>
<td>341</td>
<td>79</td>
</tr>
<tr>
<td>Dating and Relationships</td>
<td>107</td>
<td>53</td>
</tr>
</tbody>
</table>

Finally, 10 different expertise topics have been selected to evaluate the answerer finding approaches. Table 6-2 lists the 10 expertise topics and some related statistics, in which the second column is the number of users who have expertise in the topic from the 2,375 users with over 100 questions collected. In the selection of testing topics, two criteria are applied: (1) The selected 10 topics should cover a wide range of different knowledge areas; (2) Each topic should have a sufficient number of testing samples (at least 10 question-user pairs have been collected for each topic). Table 6-2 indicates that the number of testing question-user pairs collected for a topic are always less than the number of users who have expertise in the topic. This is because a user usually has knowledge of a range of different topics and it is very likely that the last 15 answered questions do not include questions in all of the users’ expertise topics. There is also a portion of questions which were not collected because of the lack of knowledge category information, i.e. the questions were not manually labelled using categorization topics.
6.4.2. Evaluation Metric and Comparative Methods

The last step is how to evaluate the answerer finding approaches using the constructed testing data. Below we describe the detailed evaluation method:

(1) Take a testing sample of a given expertise topic;

(2) Randomly select $N$ users who do not have expertise in the topic from all the 24,285 users in the dataset (based on their specified expertise topics), and merge the testing user with the $N$ users to form an answerer candidates pool;

(3) Build the testing user’s expertise model according to the answerer finding approach to be evaluated (note that the testing question is taken from the testing user’s answering history), and build other users’ expertise models using the UAH based approach;

(4) Rank the candidates based on the calculated similarity between the testing question and user expertise model.

Thus, the higher the ground truth user for a testing sample is ranked, the better the applied answerer finding approach performs. In the construction of a candidate pool for a testing sample (step 2), only users who do not have expertise in the current topic were used. It tries to reduce the effect of noise from similar users on the experimental results. If many users with expertise in the same topic are included in the candidate pool, those users could be highly ranked too in the result list. It should be noted that these similar users did not answer the testing question (i.e. they were not selected as ground truth answerers) but it is very likely that they are capable of answering the question (qualified answerers). Meanwhile, the selected ground truth user is also expected to be ranked at the top of the list. However, there is no effective way to distinguish the importance of those users and the ground truth user. In this case, it is possible that the ground truth user keeps being ranked behind the similar users. This would impact our observation on the effectiveness of different answerer finding approaches. Therefore, testing samples are selected by expertise topics and other users in the candidate pool are selected against the given topic.

In step 3, among the users in a candidate pool, the ground truth user’s expertise model is built using the applied AF approach (either the UAH based approach or the UET based approach), while that of all other users are built using the UAH based approach. The UAH based approach is first applied to all the users to observe its performance, i.e. the ranking of the ground truth user. Then, it is assumed that the ground truth user is a cold start user,
i.e. no answering history, but the user’s expertise topics are provided. In this case, the UET based approach is applied to build the ground truth user’s expertise model. It is expected that this approach can rank the ground truth user at a position as high as the UAH based approach does.

Specifically, the two types of AF approaches are implemented in the experiments as follows:

(1) **User answering history based approach**: Namely, the TF-IDF algorithm based approach as described in Section 6.2.1;

(2) **User expertise topic based approach**: To better understand the performance of the UET based approach, we experiment with its three variations. They are implemented following the steps described in Section 6.2.2 but with different settings:

- **Random expertise topics**: Randomly select $K$ expertise topics from all the 149 topics and then for each topic randomly select $N/K$ questions categorised in that topic from the question pool to simulate the user’s answering history, where $K$ is an integer randomly generated between 1 and the maximum number of expertise topics (obtained from the Quora profiles of all the users in the dataset);

- **Specified expertise topics**: Of the expertise topics specified in the user’s Quora profile, for each topic randomly select $N/K’$ questions categorised in the topic from the question pool to simulate the user’s answering history, where $K’$ is the total number of expertise topics specified in the user’s profile (only the 149 topics are considered).

- **Inferred expertise topics**: Similar to the previous variation, the only difference is that the user’s expertise topics are inferred from the user’s Twitter content in this approach (the inference approach proposed in Chapter 5 is applied for inference).

Finally, the ranking position of the ground truth user in the result list is used as the evaluation metric in the experiments. For example, if the user is ranked at the first place in the result list, the metric value is assigned as 1. The smaller the obtained metric value is, the better the answerer finding approach performs. To further alleviate the effect of possible bias in the selection of candidate users, for each testing sample, step 2 to step 4 are repeated $M$ times, i.e. randomly generate $M$ different candidate pools, and the metric
value for the testing sample is assigned as the averaged value over \( M \) candidate pools for that sample. Then, the metric values of all the samples for a topic are averaged as the final metric value for that topic. By doing so, different answerer finding approaches can be evaluated by expertise topics.

6.5. Experiments and Results

In the experiments, we set \( N = 100 \), i.e. 100 candidate users for each testing question, and \( M = 5 \), i.e. we repeat the test 5 times for each testing question. Figures 6-4 – 6-13 show the experimental results of each of the 10 testing expertise topics. It should be noted the variance among \( M \) runs of experiments for more than 95% of the testing samples is not significant (smaller than 5%).

![ Figure 6-4: Ranking accuracy of different answerer finding approaches for the topic “sales” ](image.png)

110
Figure 6-5: Ranking accuracy of different answerer finding approaches for the topic “machine learning”

Figure 6-6: Ranking accuracy of different answerer finding approaches for the topic “video games”
Figure 6-7: Ranking accuracy of different answerer finding approaches for the topic “music”

Figure 6-8: Ranking accuracy of different answerer finding approaches for the topic “recruiting”
Figure 6-9: Ranking accuracy of different answerer finding approaches for the topic “photography”

Figure 6-10: Ranking accuracy of different answerer finding approaches for the topic “philosophy”
Figure 6-11: Ranking accuracy of different answerer finding approaches for the topic "islam"

Figure 6-12: Ranking accuracy of different answerer finding approaches for the topic "startups"
6.5.1. Performance of the User Answering History Based Approach

We first vary the number of the user’s answered questions used in the range [1, 100] for the UAH based approach to observe the change of its ranking accuracy. As shown in Figure 6-13, we can see that its ranking accuracy improves as the number of questions used increases for the topic “dating and relationships”. Experimental results also show that overall, when all the questions are used, most of the ground truth answerers have been ranked in the top 5 positions. It explains that a user’s answering history can well reflect the user’s expertise and the applied UAH based approach is effective for finding the right answerers, especially when the user has sufficient previously answered questions which can be considered. It is also noted that the improvement rate of the approach slows as the number of questions used for user expertise modelling increases. This indicates that a user’s expertise on a topic can be well represented by a certain number of his/her answered questions on that topic, and considering more questions over that point will not largely influence the performance of the UAH based approach. In addition, the UAH based approach performs consistently well on each of the testing topics. It empirically demonstrates that the effectiveness of this approach is domain-agnostic.
6.5.2. Effectiveness of the Specified User Expertise Topics

The experimental results in Figures 6-4 – 6-13 show that the UET based approach using the user specified expertise topics performs as well as that of the UAH based approach (at its best ranking accuracy). In some cases, the former even outperforms the latter. For example, as shown in Figure 6-7, most of the ground truth users for all the testing questions of the topic “music” have been ranked in the top 2 positions by UET, while the UAH based approach only ranked most of them in the top 4 positions. The experimental results reveal two pieces of information: First, the user’s expertise topics are a useful information source for user expertise modelling and they can be effectively exploited to find the right answerers of new questions; Second, the user’s specified expertise topics can be even more useful in modelling the user’s expertise information, when compared with the user’s answering history. This may be due to the user answering questions that are outside his/her identified knowledge areas, which will introduce noise into user expertise modelling. In this case, extra effort may be required to select appropriate questions from the user’s answering history. In comparison, expertise topics manually specified by a user generally can represent the user’s knowledge areas in a relatively accurate manner. Similarly, the UET based approach using the specified topics performs consistently well on different expertise topics, meaning this approach is also not topic-sensitive.

6.5.3. Effectiveness of the Inferred User Expertise Topics

While the effectiveness of the user’s specified expertise topics for answerer finding is clear, it is also common that users do not provide their expertise topics on their profiles, or that the provided expertise topics are incomplete. In this case, we are trying to investigate to what extent the expertise topics inferred from the user’s social content, e.g. tweets, can be used for answerer finding on Quora. It is shown in the Figures 6-4 – 6-13 that in most cases, the ground truth users have been ranked in the top 10 positions by the UET based approach using the inferred user expertise topics. This performance is inferior to that of the approach using the user’s specified expertise topics, but greatly outperforms the random expertise topics based approach and the UAH based approach for users with less than about 40 answered questions. This suggests that the user expertise topics inferred from social media are effective in user expertise modelling, especially for those who have not yet answered a significant number of questions and have not explicitly provided their expertise topics, i.e. cold-start users. For those users, their inferred expertise topics
become very valuable and enable their inclusion in the pool of potential answerers, which can largely extend the search scope and increase the chance of finding more appropriate answerers to newly posted questions.

However, the inference accuracy for the user’s expertise topics from social media may restrict its performance in answerer finding, which is why it received an inferior ranking accuracy when compared with the approach using the user’s manually specified topics. It is also observed that the approach using the inferred expertise topics did not perform consistently well on the testing questions of different topics. For example, it achieved a ranking accuracy as good as that of the approach using the user’s specified topics for testing questions on the topic “recruiting”, while the latter greatly outperforms the former for the topic “photography”. This means the inferred expertise topics of part of the testing users are different from what the user specified, which could be caused by two possible reasons: 1) The inference approach incorrectly predicted the user’s expertise on a topic; 2) The user wrongly provided her/his expertise information on a topic. However, in the experiment of this research it is reasonable to assume that most users correctly specified their expertise on the selected topics in their profiles, which can be supported by the fact that the approach using the user’s specified topics outperforms the approach using the inferred topics on most of the tested topics (i.e. nine out of ten). So, the uncertainty in terms of the performance of the inferred topics based approach could be mainly due to the varying inference accuracy on different topics from social media. Factors, such as limited training samples for a topic and higher modelling complexity for a topic, could cause this difference in user expertise inference. Therefore, one of the main solutions to this issue is to improve the inference accuracy on particular expertise topics, which will be one of the main future focuses of this PhD research.

6.6. Chapter Summary

This chapter investigated the practical significance of the user expertise modelling approaches proposed in this PhD research. It aimed to exploit the inferred user expertise from their SNSs to facilitate the answerer finding service in the popular CQA site Quora and demonstrate the usefulness of this inferred information in a real-world application scenario. To evaluate the answerer finding approach using the inferred expertise, experimental data was harvested from Quora that includes over 20,000 Quora users and over half a million questions answered by these users. In the experiments, users who actually answered a question were taken as the ground truth answerer to that question,
and questions in ten different expertise topics were eventually harvested in the dataset to evaluate different AF approaches. Experimental results showed that the AF approach using the inferred user expertise topics outperforms the approach using the randomly generated expertise topics and the user answering history based approach using less than 40 answered questions for expertise modelling. This result suggests that the proposed user expertise modelling approaches using social media data in this PhD research are useful to model the expertise of users who did not answer sufficient questions and did not explicitly provide their expertise topics, i.e. cold-start users. Thus, they can enable the inclusion of cold-start users in the pool of potential answerers, which can largely extend the search scope and increase the chance of finding more appropriate answerers to newly posted questions.
7. Conclusion

This chapter discusses the findings from the experiments presented in this thesis, and draws the final conclusions for this PhD research. Section 7.1 reviews the research objectives of this PhD research (set out in Chapter 1) and discusses to what extent these objectives have been achieved. Section 7.2 revisits the contributions of this PhD research. Section 7.3 discusses possible future research directions on the basis of research outcomes achieved in this PhD thesis. To end, final remarks are presented in Section 7.4.

7.1. Research Objectives vs Research Achievements

The question which this research initially set out to answer was:

*To what extent can a user’s content, actions and connections on social networking sites be exploited by novel user modelling approaches to infer their expertise?*

To assess whether the user expertise inferred by the proposed approaches is of benefit to information services, a further question was also addressed:

*Can the inferred user expertise be used to facilitate services such as answerer finding on a community question answering site?*

To address the two research questions, three specific research objectives were posed (see Section 1.3). This section first reviews the three objectives and then analyses in detail the extent to which each research objective was achieved.

7.1.1. Research Objectives Recap

- **Objective 1**: Investigate the possibility of modelling the expertise of cold start users in SNSs; propose corresponding modelling approaches to infer their expertise information from user-generated content such as connections and actions, and experimentally evaluate the effectiveness of the proposed approach;

- **Objective 2**: Propose effective user modelling approaches to infer a user’s topical expertise by exploiting their various activities in a SNS, i.e. non-cold start users, and experimentally evaluate the effectiveness of the proposed approach;

- **Objective 3**: Apply the user’s inferred expertise information from their social content to a community question answering site for answerer finding for new questions, and
experimentally evaluate the effectiveness of the answerer finding approach using inferred expertise information (via the approaches proposed in this PhD research) for answerer finding.

7.1.2. Evaluation of Research Achievement

7.1.2.1. Objective 1: Expertise Inference of Cold Start Users

This research objective was achieved through the design of the Language and Social Relation-based Factor Graph Model (LSR-FGM) for user language inference using the user’s social profiles, and the related evaluation experiments on a LinkedIn dataset, which are described in Chapter 3.

It first identified that user profiles are an important source of first-hand information about a new user in SNSs, because users are often asked to provide some basic personal information like education and work experience when registering, such as in Facebook and LinkedIn. Also, it was observed that the experiences described in a user’s profile could imply what languages they may understand. Thus, in Chapter 3, it was proposed to exploit the static profiles of new users in a SNS to infer their language expertise. To overcome the problems of data incompleteness and similarities between inference tasks in modelling the problem, a language and social relation-based factor graph model was proposed, which collectively considers three key factors for language expertise inference: textual attributes of user profiles, relations between languages and social relations between users. In LSR-FGM, attributes in a user’s profile provide fundamental evidence for user language inference. Taking the language relation into consideration can help to overcome difficulties in inferring the user’s expertise in multiple languages, and the use of social relations can help to alleviate the problem of incompleteness of user profiles.

To evaluate the effectiveness of the proposed LSR-FGM, real-world LinkedIn profiles were used to build a benchmark dataset for the language inference task. In the experiments, five languages were considered as the target inference languages of the task. Experiments compared the inference performance of LSR-FGM with that of several alternative modelling approaches, and experimental results showed that LSR-FGM consistently achieves the best inference performance on all five testing languages. Additional experiments were conducted to examine the contribution of the three factors in LSR-FGM, and experimental results confirmed that each factor makes a stand-alone contribution to the process of expertise inference.
7.1.2.2. Objective 2: Expertise Inference of Active Users

This research objective was achieved through the design of the Sentiment-weighted and Topic Relation-regularized Learning (SeTRL) model for user topical expertise inference using the user’s posted tweets on Twitter, as well as a learning model that discriminately exploits multiple types of user data on Twitter for user topical expertise inference, and related evaluation experiments on a Twitter dataset, which are described in Chapter 4 and Chapter 5.

Due to the public availability of user data on Twitter, this PhD research focused on the Twitter platform for the study of expertise inference of active SNS users. Firstly, it looked at the question of inferring a user’s topical expertise based on the content (i.e. tweets) that the user had previously posted (Chapter 4). The major challenge faced by this research is the noise inherent in social content, i.e. users can publish content on any topic in the social media environment so frequent mentions of terms about a topic in a user’s authored content does not necessarily indicate that this user has a good level of knowledge of that topic. To address the challenge, a sentiment-weighted and topic relation-regularized learning model was proposed that exploited two types of prior knowledge to better model the problem: the sentiment intensity of tweets and a dependency relation between expertise topics. In the SeTRL model, the sentiment intensity of a tweet is used to evaluate its importance in inferring a user’s expertise, which is based on the idea that if a person can forcefully express their opinion on a topic, it is more likely that the person has strong knowledge of that topic. As there are intrinsic relations between the topics of expertise of an individual (i.e. the existence of the probability that a user who knows topic A is more likely to know topic B than a user who does not know topic A), this relatedness between expertise topics is exploited in the SeTRL model through regularization to improve the inference performance.

To evaluate the proposed SeTRL model for user expertise inference, it was necessary to construct a benchmark dataset in which the expertise topics of a large number of Twitter users were identified. This PhD research utilised the account connections between the popular CQA site Quora and Twitter, and built a dataset consisting of 149 expertise topics and over 10,000 Twitter users. Experiments conducted on the dataset first verified that the tweet sentiment-based weighting scheme is superior to several other alternative schemes in evaluating the importance of user features for expertise inference. Then, experiments compared the inference performance of the SeTRL model with that of
several alternative modelling approaches, and results showed that SeTRL achieved the best overall performance on all the tested expertise topics. Additionally, experiments examined the contribution of the prior knowledge of topic relatedness in the model and results demonstrated that it makes a stand-alone contribution in the process of user expertise inference.

Although SeTRL was shown to be effective in inferring a user’s expertise by exploiting the user’s posted tweets, many users on Twitter rarely post tweets. As mentioned in Chapter 1, it was reported that about 44% of all registered users on Twitter have never posted a tweet. In this case, the proposed SeTRL model will fail for user expertise inference. Therefore, this PhD research (Chapter 5) further studied the problem of user expertise inference on Twitter, with a focus on the exploitation of various user activities, i.e. multiple types of user data on Twitter. On the basis of the SeTRL model, a multi-Data and Topic relatedness Combined (D^nTCom) learning model was proposed that takes multiple types of user data as input and aims to ensure the inference effectiveness regardless of the availability of some types of user data. In the D^nTCom learning model, it also considered the inference consistency of different types of user data in the process of inference and assumed that the expertise information of an individual reflected by her/his different types of user data should be similar.

The same Twitter dataset which was constructed in Chapter 4 was re-used to evaluate the proposed D^nTCom learning model. This PhD research considered four different types of user data on Twitter to infer the user’s expertise, i.e. tweets, friends, followers and lists of a Twitter user. Therefore, the four types of data were harvested for each user in the dataset for the evaluation experiments. In the experiments, the inference performance of each type of user data was first tested separately, with different feature spaces and different feature weighting schemes, using state-of-the-art inference approaches. It demonstrated that each type of user data can be exploited to effectively infer the user’s expertise, and that this process also helped to select an appropriate feature space for the problem of user expertise inference from multiple types of user data (i.e. unigram features) as well as the corresponding weighting scheme. Then, experiments were conducted to examine the performance of different methods using the combination of all the four types of user data for expertise inference. Experimental results showed that the proposed D^nTCom significantly outperforms other alternative inference approaches (including the SeTRL model). Finally, experiments compared the performance of D^nTCom with various
combinations of the four types of user data. The results showed that D⁰TCom using all four types of data performs the strongest, when compared with all other different combinations. The results also demonstrated that tweet and friend data are more effective for user expertise inference, when compared with followers and lists.

7.1.2.3. Objective 3: Application of the Inferred Expertise Information

This research objective was achieved through the design of experiments to investigate the usefulness of the inferred user expertise information using the proposed expertise modelling approaches in a real-world application scenario, namely a community question answering site. These experiments are described in Chapter 6.

To study the practical significance of the proposed user expertise modelling approaches in real-world applications, this PhD research focused on the popular CQA site, Quora, and aimed to improve its service through the use of the proposed expertise modelling approaches. In order to deliver a competent CQA service, efficiently and accurately finding users who have the ability to answer newly posted questions is one of its most important challenges. This process is named Answerer Finding (AF) in this thesis. However, as discussed in Chapter 6, a large proportion of users on CQA sites, including Quora, have answered few, if any, questions, i.e. cold start users. Therefore, the expertise of those users cannot be modelled from their answering history. This fact greatly limits the effectiveness of traditional user answering history based approaches for AF. To attempt to alleviate this problem, this PhD research proposed to model the cold start users’ expertise information from their Twitter activities using the proposed expertise modelling approaches (Chapters 4 and 5), and take them into consideration in answerer finding, with the aim of increasing the chance of finding appropriate answerers for new questions.

Many Quora users provide their Twitter account in their profiles, which allows us to link Quora users with their Twitter accounts and build up a benchmark dataset for the evaluation experiments. The dataset constructed by this research includes over half a million questions, answered by over 20,000 Quora users. In the experiments, testing questions were selected by expertise topics, i.e. questions belong to a certain knowledge category (10 different expertise topics were used). Users who in reality actually answered a question were considered to be the target answerers for that question. Experiments were conducted to examine the effectiveness of the AF approach using the inferred expertise information, and results showed that it outperforms the approach using the randomly
generated user expertise topics and the user answering history based approach using less than \( \sim 40 \) answered questions for expertise modelling. This result suggests that the proposed user expertise modelling approaches using social network data in this PhD research are useful to model the expertise of users who have not yet answered a significant number of questions and have not explicitly provided their expertise topics, i.e. cold-start users. They can enable the inclusion of cold-start users in the pool of potential answerers, which can significantly extend the search scope and increase the chance of finding more appropriate answerers to newly posted questions.

7.2. Contributions

This section briefly revisits the contributions from the research of this PhD thesis, which were initially presented in Section 1.4.

The research of this thesis yielded a major contribution and a minor contribution. The major contribution is the three user modelling approaches that infer a user’s expertise information by exploiting their various social content on SNSs: (1) Language and Social Relation-based Factor Graph Model (LSR-FGM) that exploits the user’s social profiles to infer their language expertise; (2) Sentiment-weighted and Topic Relation-regularized Learning (SeTRL) model that exploits the user’s posted tweets on Twitter to infer their topical expertise and (3) multi-Data and Topic relatedness Combined (D\(^n\)TCom) learning model that exploits multiple types of user data on Twitter to better infer the user’s topical expertise.

The LSR-FGM is a learning model that enables us to predict the languages that cold start users comprehend through the use of their static social profiles. LSR-FGM is novel in that, it is the first attempt to acquire the online users’ language information via an analysis of their social experiences, instead of their writing or reading history. The LSR-FGM advances the state of the art, because, apart from the textual attributes of social profiles, it also incorporates two advanced structural factors in profiles, i.e. dependency relations between languages and social relations between users, to enhance the language prediction accuracy. LSR-FGM offers the community an effective approach for language prediction of cold start users via the exploitation of limited information about them, i.e. their static social profiles. Experimentation conducted on the LSR-FGM demonstrates the existence of the correlation between the user’s language expertise and the proposed structural
factors. This contributes a basis for the development of more advanced approaches in this area.

Aiming to infer the topical expertise of active Twitter users, SeTRL is a learning model that allows us to predict the topical expertise of Twitter users through the use of their previously posted tweets. SeTRL advances the state of the art as it uses a novel approach to evaluate the importance of user features (extracted from their tweets) in expertise inference, i.e. the tweet sentiment analysis based approach, which has been shown to be superior to the traditional word frequency based approaches. Additionally, SeTRL also incorporates prior knowledge of relations between expertise topics into the process of inference in order to deliver better prediction performance. SeTRL offers the community an effective approach to infer Twitter users’ topical expertise based on their posted tweets. The discovered impact of tweet sentiment and/or topic relation on expertise inference contributes a research foundation for future study in this area.

Based on SeTRL, the multi-Data and Topic relatedness Combined (D^nTCom) learning model is proposed that allows us to predict the topical expertise of Twitter users through the collective use of their various activities on Twitter (including previously posted tweets). The D^nTCom advances the state of the art (including SeTRL) as it can deliver effective inference as long as there are some types of user data available. It takes multiple types of user data as input for expertise inference and overcomes the problem that many Twitter users tend to rarely conduct certain types of activities, e.g. they never post any tweets, which will lead to the failure of approaches only relying on a single type of user activity. In addition, the D^nTCom learning model also takes into consideration the inference consistency of different types of user data in order to improve inference performance. D^nTCom offers the community an effective approach to infer the topical expertise of a larger amount of Twitter users (when compared with SeTRL), where as long as there are some types of activities conducted by the user on Twitter, inference can be performed.

The minor contribution is the design and implementation of experiments that investigate the usefulness of the proposed expertise modelling approaches in a real-world application. This research takes the popular CQA site Quora as the target application and aims to improve answerer finding on the platform through the use of the proposed expertise modelling approaches. Experimental data is harvested to simulate the application scenario, where cold start users are not considered by the traditional answering history
based approaches in finding potential answerers for new questions, even though many of them have linked their Quora accounts to their social media accounts. An answerer finding approach is proposed that models the cold start user’s expertise information based on their Twitter activities (using the proposed modelling approaches in this PhD research) and then matches them with the newly posted questions on Quora. Evaluation methods are designed to test the performance of the new answerer finding approach and results show that it outperforms the answering history based approach when less than ~40 answered questions are considered for user expertise modelling. This means the proposed user expertise modelling approaches are useful to model the expertise of cold start users on CQA sites and enable the inclusion of them in the pool of potential answerers. The evaluation experiments not only further show the significance of the proposed user modelling approaches, but also demonstrate a specific example for their use in more application scenarios.

7.3. Further Work
This section discusses potential short-term and long-term further work that could be undertaken for the research in this PhD thesis.

7.3.1. Short-Term Future Work
Following on the research carried out in this PhD thesis, this sub-section discusses potential work that could be undertaken in the short-term future. The LSR-FGM considers the language expertise inference of SNS users from their social profiles. In terms of the future work on the basis of the proposed LSR-FGM, one possible route is to exploit more external resources to enhance the user profiles. Social profiles initially provided by the users are usually incomplete, thus, it is important to associate them with more related information that could help language inference, such as the locations they have worked. Another possible route is the design of more advanced features that may be related to the user’s language expertise. Although the sentiment analysis based approach has been shown as an effective feature weighting scheme for user expertise inference, future work could also involve exploring other weighting schemes that can better represent the user’s expertise, such as the use of the user’s hedging language [Ma13]. In terms of the general topic of expertise inference of cold start users in SNSs, future work may still focus on the use of the user’s social profiles, but with the aim to infer other expertise information about the user. For example, the studying and working experiences stated in the user’s profile
could also indicate what professional skills the user may have. Apart from the use of static profiles, other limited interactions cold start users have in a SNS should also be explored, such as the list of users the user follows/connects to. In this case, studies are expected to identify their connections with the user’s expertise information.

Another focus of this PhD research is the topical expertise inference of active Twitter users. On the top of the two learning models proposed, one possible future route is to design more effective approaches of constructing relations between expertise topics. It was shown that the relations between expertise topics play an important role in the inference of the user’s topical expertise, however, the approach applied in the SeTRL model for constructing the relation is relatively simple, i.e. based on their co-occurrence frequencies. Thus, it may be possible to exploit more resources, or conduct user studies, to obtain the relations that can better represent the intrinsic connections between expertise topics of an individual. Another possible route of the future work is to investigate the impact of inference models on different expertise topics. The design of the learning models in this PhD thesis targets for better overall inference performance on all the expertise topics, so its performance on different topics may be inconsistent. In application scenarios, they may need to improve its performance on particular topic(s), while maintaining its overall performance. Thus, there is a need for the design of topic-aware learning models for expertise inference.

The last focus of this PhD research is the application of the inferred user expertise using the proposed modelling approaches. As there is a lack of multilingual content on Quora, the inferred language expertise was not applied in the case study research. Thus, the construction of related datasets with multilingual content or the study on the application of the inferred language information in other potential application scenarios will be one possible route of future work. In addition, the case study research validated the usefulness of the inferred user expertise and did not explore to what extent the inferred expertise can facilitate the answerer finding service on a CQA site. Thus, another possible route of future work is to design approaches that can more effectively utilise the inferred user expertise on different application scenarios. Furthermore, it is also noted that the inactivity of cold start users in a CQA site may be part of the classic online community lurker behavior. In this case, effective modelling of these users will not help to encourage them answer new questions. To address this type of problem, the enhanced user profile from their SNS content could be used to infer motivational factors, such as targeted prizes
and direct attention from the community moderators, to help engage them in question answering.

7.3.2. Long-Term Future Work
This sub-section discusses potential research directions this PhD research could lead to in the long-term future. This PhD research discovered a variety of user expertise-related factors, such as tweet sentiment and topic relatedness, and incorporated them in novel user expertise modelling approaches for effective user expertise inference in the SNS setting. It is a straightforward future route to keep exploring other factors that are associated with the user’s expertise information and optimizing expertise modelling approaches accordingly. However, the limitation in accessing the SNS data would prevent the development of this direction of research. It was noted that the exportation of this PhD research was limited to the public LinkedIn profiles and public data of the Twitter users. Therefore, to continue on this line of research in the long term, expanding access or gaining full access to data in SNS platforms, such as Twitter or Facebook, is required.

This PhD research focused on the utilization of the publicly available user data and did not involve privacy related issues. However, as indicated in this thesis, the proposed expertise modelling approaches can be readily applied to other SNS platforms, which would bring in concerns on the use of the privacy-sensitive user data in platforms such as Facebook. This will lead to a new line of research, i.e. privacy-aware user expertise modelling in SNSs. Future research could involve designing effective user expertise modelling approaches under the context of controlled user data access or effective privacy protection strategies in user expertise modelling.

Researchers in academia often suffer the problem of the lack of real-world experimental data. They are usually at a disadvantage when performing research that relies on real-world user data, compared with the research teams within that platform, e.g. Twitter and Facebook. Under this background, this leads this PhD research to another line of research, which focuses on user expertise modelling for targeted groups of users through the use of their SNS content and local documents, such as employees within large multinational companies or researchers within large research centres. Academic researchers are usually affiliated with certain research groups or in collaboration with some industry companies. This allows them to have access to the data of certain groups of people, which could include their local documents, e.g. project files and publications, and SNS content, e.g. Facebook updates. Although extensive research has been conducted to study the problem
of user expertise modelling within an organization by using their local documents (reviewed in Section 2.2), the research question: *to what extent can the user’s social content and local documents within an organization be jointly exploited to model their expertise*, remains unanswered. Therefore, future research could focus on the design of effective expertise modelling approaches through the combinational use of the user’s various social content and local documents. Furthermore, based on the constructed user expertise model, future research could also involve the design of effective user search approaches or user recommendation approaches, in order to serve targeted applications within the organization, such as personalized talent search, task match.

### 7.4. Final Remarks

It is hoped by the author of this thesis that the proposed expertise inference approaches, that support the modelling of the user expertise information in SNSs, will be of benefit to both the research community and commercial applications. The proposed LSR-FGM will be useful for researchers who wish to model the language information of SNS users and those who wish to utilise the user’s social profiles for other inference tasks. The proposed SeTRL and DºTCom for topical expertise inference on Twitter will be useful for researchers who wish to model online users’ topical expertise by exploiting various social activities. Researchers could continue to contribute to new/advanced inference approaches in the future, with a target of the exploitation of various other social data (e.g. the involvement of multimedia data) or the inference of different aspects of user expertise (e.g. expertise related to the targeted tasks/projects).

It is also hoped that the proposed inference approaches will be of benefit to the providers of SNSs and those Web applications that linked to their users’ SNS accounts through the adoption of these proposed approaches. The case study experiments in this thesis have shown the benefits that these approaches can bring to a CQA service. As a result, more Web applications or SNS providers themselves may employ these approaches to gain a better understanding of their users, which would allow them to enhance their current services or even assemble new functionalities.
References


[Li10a] Liu, Mingrong, Yicen Liu, and Qing Yang. "Predicting best answerers for new questions in community question answering." In Proceedings of


[We10] Weng, Jianshu, Ee-Peng Lim, Jing Jiang, and Qi He. "Twitterrank: finding topic-sensitive influential twitterers." In Proceedings of the third ACM

[We16] Wei, Wei, Gao Cong, Chunyan Miao, Feida Zhu, and Guohui Li. "Learning to find topic experts in Twitter via different relations." IEEE Transactions on Knowledge and Data Engineering. 28(7): 1764-1778, 2016.


[Ya14b] Yang, Jie, Ke Tao, Alessandro Bozzon, and Geert-Jan Houben. "Sparrows and owls: Characterisation of expert behaviour in stackoverflow." In the


