Privacy Preserving Neighbourhood-based Collaborative Filtering Recommendation Algorithms

Zhigang Lu

School of Computer Science

The University of Adelaide

A thesis submitted for the degree of

Master of Philosophy

August, 2015

To my loving parents and fiancée …

Declaration

I , Zhigang Lu, certify that this work contains no material which has been accepted for the award of any other degree or diploma in my name, in any university or other tertiary institution and, to the best of my knowledge and belief, contains no material previously published or written by another person, except where due reference has been made in the text. In addition, I certify that no part of this work will, in the future, be used in a submission in my name, for any other degree or diploma in any university or other tertiary institution without the prior approval of the University of Adelaide and where applicable, any partner institution responsible for the joint-award of this degree.

I give consent to this copy of my thesis when deposited in the University Library, being made available for loan and photocopying, subject to the provisions of the Copyright Act 1968. The author acknowledges that copyright of published works contained within this thesis resides with the copyright holder(s) of those works.

I also give permission for the digital version of my thesis to be made available on the web, via the University's digital research repository, the Library Search and also through web search engines, unless permission has been granted by the University to restrict access for a period of time.

Signature: _____ Date: _____

Acknowledgements

This MPhil study was a truly challenge and amazing life experience for me, it would not be possible without the support and guidance that I received from many people.

First and foremost, I would like to express my special appreciation and thanks to my principal supervisor, Prof. Hong Shen, for his great and patient supervision over the past two years. Under Prof. Shen's supervision, I have learned how to write research paper, how to make academic presentation, how to be a teaching assistant and so on, but the most important thing I learned from Prof. Shen is how to think as a researcher, such as the ability to find unsolved research problems and address them in a novel way. Moreover, Prof. Shen is the one who leads me to the filed of research, encourages me to continue my research career in coming years.

I would also like to express my gratitude to my supervisor, Dr. Nickolas Falkner, for his help, support in my research.

I own my parents and fiancée a lot, for their unconditional love, support and encourage. Especially, without the financial support from my parents, I even do not have the chance to study at the University of Adelaide.

I would like to thank some of the university staff, my colleagues and friends who provide the selfless help during my two years MPhil study: Ms. Jo Rogers, Ms. Julie Mayo, Ms. Sharyn Liersch, Ms. Lenka Hill, Ms. Sue Fiedler, Dr. Yuval Yarom, Ms. Alison Black, Ms. Rosie Wilkes, Dr. Crusher Wong, Dr. Kewen Liao, Yongrui, Dr. Xiaoqiang Qiao, Yihong, Scott, Ali, Javier, Zheng, Jack, Dr. Yingpeng Sang, Lingjing, Dr. Sergey Polyakovskiy, Dr. Jian Kang, Dr. Mingkui Tan, Wenjie, Dung.

Finally, I thank the School of Computer Science for financially supporting my conference travels. Thanks also go to the staff in Adelaide Graduate Centre and the anonymous examiners for their efforts on this thesis.

Preface

During my two years MPhil study at the University of Adelaide, I have produced one journal article and two conference papers related to this thesis. The thesis is based on the content presented in the following papers.

Journal Publication:

- Zhigang Lu and Hong Shen. (2015). An accuracy-assured privacy-preserving recommender system for Internet commerce. *Computer Science and Information Systems*. (minor revision)

Conference Publications:

- Zhigang Lu and Hong Shen. (2015). A security-assured accuracy-maximised privacy preserving collaborative filtering recommendation algorithm. In *Proceedings of the 19th International Database Engineering & Applications Symposium*, IDEAS'15, pages 72-80, New York, NY, USA. ACM.
- Zhigang Lu and Hong Shen. (2015). A fast algorithm to build new users similarity list in neighbourhood-based collaborative filtering. to appear in *Proceedings of the 16th International Conference on Parallel and Distributed Computing, Applications and Technologies*, PDCAT'15, Jeju, Korea.

Abstract

Recommender systems, which recommend users the potentially preferred items by aggregating similar interest neighbours' history data, show an increasing importance in various Internet applications. As a well-known method in recommender systems, neighbourhood-based collaborative filtering has received considerable attention recently because of its easy implementation and high recommendation accuracy. However, the risks of revealing customers' privacy during the process of filtering have attracted noticeable public concern. Specifically, the *k*NN attack discloses the target user's sensitive information by creating *k* fake nearest neighbours by non-sensitive information. Among the current solutions against the *k*NN attack, probabilistic methods showed a powerful privacy preserving effect. However, the existing probabilistic methods neither guarantee enough prediction accuracy due to the global randomness, nor provide assured security enforcement against the *k*NN attack.

To overcome the problem of recommendation accuracy loss, we propose a novel approach called Partitioned Probabilistic Neighbour Selection. In this thesis, we define the sum of *k* neighbours' similarity as the accuracy metric α, the number of user partitions, across which we select the *k* neighbours, as the security metric β. We consider two versions of the Partitioned Probabilistic Neighbour Selection schemes. Firstly, to ensure a required prediction accuracy while maintaining high security against the *k*NN attack, we propose an accuracy-assured Partitioned Probabilistic Neighbour Selection algorithm. We select neighbours from each exclusive partition of size *k* with a decreasing probability. Theoretical and experimental analysis show that to provide an accuracy-assured recommendation, our method yields a suitable trade-off between the recommendation accuracy and system security. Secondly, to ensure a required security guarantee while achieving the optimal prediction accuracy against the *k*NN attack, we propose a security-assured accuracy-maximised Partitioned Probabilistic Neighbour Selection algorithm. We select neighbours from each partition with exponential differential privacy to reduce the magnitude of noise. Theoretical and experimental analyses show that to achieve the same security guarantee against the *k*NN attack, our approach ensures an optimal prediction accuracy.

In addition, as the core of neighbourhood-based CF, the task of dynamically maintaining users' similarity list is challenged by the cold-start problem and the scalability problem. Recently, several methods have been proposed for solving the two problems. However, these methods require mn steps to compute the similarity list against the kNN attack, where n and m are number of users and items respectively. Observing that the k new users from the kNN attack, with enough recommendation data, share the same rating list, we present a faster algorithm, TwinSearch, to avoid computing and sorting the similarity list for each new user repeatedly to save the time complexity. The computation cost of our algorithm is $\frac{1}{125}$ of the existing methods. Both theoretical and experimental results show that the TwinSearch Algorithm achieves better running time than the traditional method.

Table of Contents

List of Figures xv

List of Tables xvii

List of Symbols xix

1 Introduction 1

1.1 Research Background 1

1.2 Overview of Contributions 3

1.3 Outline 4

2 Background 7

2.1 Privacy Preserving CF Recommendation Algorithms 7

2.1.1 Traditional Privacy Preserving Recommender Systems 7

2.1.2 Differential Privacy Recommender Systems 9

2.2 Reducing Computational Cost 11

3 Preliminaries 13

3.1 k Nearest Neighbour Collaborative Filtering 13

3.2 Differential Privacy 15

3.3 Wallenius' Non-central Hyper-geometric Distribution 16

3.4 A Generalised Privacy Attack for Recommender Systems 18

3.4.1 The k Nearest Neighbour Attack 18

3.4.2 β-k Nearest Neighbours Attack 18

3.5 Performance Metrics 19

3.5.1 Accuracy 19

3.5.2 Security 20

4 An Accuracy-Assured Privacy-Preserving CF Algorithm 21
4.1 Introduction . . . 21
4.2 Partitioned Probabilistic Neighbour Selection Algorithm . . . 22
4.3 Theoretical Analysis . . . 24
4.3.1 Accuracy Analysis . . . 24
4.3.2 Security Analysis . . . 26
4.3.3 Analysis Results . . . 28
4.3.4 A representative Example . . . 28
4.4 Experimental Evaluation . . . 29
4.4.1 Dataset and Experimental Settings . . . 29
4.4.2 Experimental Results . . . 30

5 A Security-Assured Accuracy-Maximised Privacy-Preserving CF Algorithm 33
5.1 Introduction . . . 33
5.2 Partitioned Probabilistic Neighbour Selection Scheme against the kNN attack 34
5.3 Partitioned Probabilistic Neighbour Selection Scheme against the β-kNN attack . . . 37
5.4 Experimental Evaluation . . . 41
5.4.1 Datasets and Experimental Settings . . . 41
5.4.2 Experimental Results . . . 42

6 A Faster Algorithm to Build New Users Similarity List in Neighbourhood-based CF 49
6.1 Introduction . . . 49
6.2 The TwinSearch Algorithm . . . 50
6.2.1 Algorithm Design . . . 50
6.2.2 Time Complexity Analysis . . . 51
6.3 Experimental Evaluation . . . 54
6.3.1 Dataset and Experimental Settings . . . 54
6.3.2 Experimental Results . . . 54

7 Conclusion 57

References 59

List of Figures

4.1 Impacts of p on accuracy 30
4.2 Impacts of p on security 30
4.3 Impacts of k on accuracy 31
4.4 Impacts of ρ on accuracy 31

5.1 Candidate list against the kNN attack 35
5.2 Candidate list in the 1-kNN attack 38
5.3 Candidate list in the i-kNN attack 39
5.4 Item-based prediction accuracy on MovieLens 42
5.5 User-based prediction accuracy on MovieLens 43
5.6 User-based prediction accuracy on Douban film 43
5.7 Prediction accuracy on MovieLens against the kNN attack 44
5.8 Impacts of k on prediction accuracy against the kNN attack on MovieLens . 44
5.9 Impacts of ε on prediction accuracy against the kNN attack on MovieLens . 45
5.10 Impacts of m on prediction accuracy against the kNN attack on MovieLens . 45
5.11 Impacts of β_0 on prediction accuracy against the β-kNN attack 46
5.12 Impacts of each x-kNN attack on prediction accuracy against the β-kNN attack 47
5.13 Impacts of m on prediction accuracy against the β-kNN attack 47
5.14 Impacts of k on prediction accuracy against the β-kNN attack 48
5.15 Impacts of ε on prediction accuracy against the β-kNN attack 48

6.1 Distribution of user's similarity list 53
6.2 Running time of User-based CF on MovieLens 55
6.3 Running time of User-based CF on Douban film 55
6.4 Running time of Item-based CF on MovieLens 56
6.5 Running time of Item-based CF on Douban film 56

List of Tables

4.1 Impacts of p on accuracy 30
4.2 Impacts of p on security 30
4.3 Impacts of k on accuracy 31
4.4 Impacts of ρ on accuracy 31

List of Symbols

$N_k(u_a)$ User u_a's k nearest neighbours set, descending order of similarity

α Accuracy metric

$\bar{r}_i$ User u_i's average rating on every rated item

β Security metric

$\mathscr{S}_a$ User u_a's similarity list

$f_\beta(i)$ The number of neighbours selected from partition No. i with the given security metric β

kNN k Nearest Neighbour

r_{is} User u_i's rating on item t_s

$sim(a,i)$ Similarity between user u_a and u_a's ith neighbour in $N_k(u_a)$

sim_{ij} Similarity between user u_i and user u_j

CF Collaborative Filtering