

Enhanced Web Log Based Recommendation by Personalized Retrieval



Xueping Peng

FACULTY OF ENGINEERING AND INFORMATION TECHNOLOGY

UNIVERSITY OF TECHNOLOGY, SYDNEY

A thesis submitted for the degree of

Doctor of Philosophy

February 2015

CERTIFICATE OF AUTHORSHIP/ORIGINALITY

I certify that the work in this thesis has not previously been submitted for a degree nor has it been submitted as part of requirements for a degree except as fully acknowledged within the text.

I also certify that the thesis has been written by me. Any help that I have received in my research work and the preparation of the thesis itself has been acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

Signature of Student

Acknowledgements

First I would like to thank my supervisors. Prof Chengqi Zhang, Prof Zhendong Niu and Dr Ling Chen who introduced me to the wonderful world of research. They not only gave me invaluable academic advice, but also helped my transition into a different culture. Prof Chengqi Zhang has been a great mentor and collaborator, being both energetic and full of ideas. Prof Zhendong Niu, who works in the School of Computer Science, Beijing Institute of Technology, guided me in terms of information retrieval and web log mining. Dr Ling Chen helped me by asking insightful questions, and giving me thoughtful comments on the thesis. I enjoyed working with them, and benefited enormously from my interactions with them.

I spent nearly two and a half years in the University of Technology, Sydney. I thank the collaborators, faculty staff, fellow students and friends in the QCIS centre, who made my graduate life a very memorable experience. In particular, I thank you if you are reading this thesis.

I thank my family. My parents richly endowed me with curiosity about the natural world. Last but not least, my deep gratitude is extended to my dear wife Suling Niu who brings me so much love and happiness. It is no exaggeration to say that she helped to make my thesis writing an enjoyable endeavor.

Contents

Contents	vi
List of Figures	xii
List of Tables	xiv
1 Introduction	1
1.1 Background	1
1.2 Research Questions	3
1.3 Research Objectives	4
1.4 Significance and Main Contributions	6
1.5 Research Methodology	7
1.6 Thesis Structure	10
1.7 Publications Related to This Thesis	12
2 Literature Review	15
2.1 Web Log Mining	15
2.1.1 Technologies	16
2.1.1.1 Data Collection	16

2.1.1.2	Data Preprocessing	17
2.1.1.3	Mining Algorithms	19
2.1.1.4	Pattern Analysis	19
2.1.2	Research and Applications	19
2.1.3	Challenges	21
2.2	Personalized Retrieval	21
2.2.1	Query Expansion	22
2.2.2	Result Processing	24
2.2.3	Challenges	24
2.3	Recommender System	25
2.3.1	Collaborative Filtering	25
2.3.2	Content-Based Approach	26
2.3.3	Hybrid Approach	27
2.3.4	Challenges	28
2.4	Summary	29
3	Query Suggestion Model Based on The Query Semantics and Click-through	
	Data	31
3.1	Introduction	31
3.2	Related Works	33
3.2.1	Query Suggestion	33
3.2.2	Bipartite Graph	33
3.3	The Proposed Method	35
3.3.1	Query Semantics for Document-based Method	35
3.3.2	Query-URL Bipartite Graph for Log-based Method	38

3.3.2.1	Construction of Query-URL Matrices	38
3.3.2.2	Matrix Factorisation and Query Relevance Computation	39
3.3.3	Integrate Multiple Suggestion Models	41
3.4	Experiments	41
3.4.1	Data Set	41
3.4.2	Evaluation Metrics	42
3.4.3	Comparison of Query Suggestion Results	42
3.4.4	Evaluation of Suggestion Results	43
3.5	Summary	44
4	Collaborative Filtering Retrieval Model Based on Local and Global Features	47
4.1	Introduction	47
4.2	Related Works	49
4.2.1	Personalized Web Search	49
4.2.2	Filtering Algorithms	50
4.2.3	Click-through Data	51
4.3	Proposed Approach	52
4.3.1	User Profile	52
4.3.1.1	Sequence Score	52
4.3.1.2	Web Page Rating	53
4.3.1.3	Preference Score	53
4.3.2	User-Based Collaborative Filtering	54
4.3.3	Personalized Search Model	55
4.4	Experiments	56
4.4.1	Data Sets	56

4.4.2	Evaluation Metrics	57
4.4.3	Ranking Methods Compared	58
4.4.4	Users Evaluation	59
4.4.5	Impact of Parameters	60
4.4.6	Personalized Search Performance	62
4.5	Summary	62
5	Web Search Recommendation Based on the Retrieval Sequence and the Browsing Features	63
5.1	Introduction	63
5.2	Related Works	65
5.2.1	User Information Collection	65
5.2.2	Web Search Recommendation	66
5.3	The Proposed Method	67
5.3.1	Definition of Terms	67
5.3.2	User Modeling	69
5.3.3	Recommendation of Resource	71
5.3.3.1	Resource Retrieval	71
5.3.3.2	Resource Filtering	71
5.4	Experiments and Analysis	73
5.4.1	Data Set	73
5.4.2	The Standard of Evaluation	73
5.4.3	Results of Experiment	74
5.5	Summary	75

6	Recommendation Based on User Interests Association Findings	77
6.1	Introduction	77
6.2	Related Works	78
6.2.1	Association Rule Mining	78
6.2.2	Maximal Frequent Itemsets	79
6.3	The Proposed Method	79
6.3.1	Basic Concepts and Assumptions	79
6.3.1.1	Basic Concepts	79
6.3.1.2	Basic Assumptions	80
6.3.2	Resources Collection and Description	81
6.3.2.1	Resources Collection	81
6.3.2.2	Resources Description	81
6.3.3	User Profile	82
6.3.3.1	Association Algorithm	82
6.3.3.2	User Modelling	83
6.3.4	Recommendation of Resources	84
6.4	Experiments and Analysis	85
6.4.1	Experiment Data Set	85
6.4.2	Evaluation Metrics	86
6.4.3	Analysis of Experiment Results	86
6.5	Summary	87
7	Conclusions and Future Research	89
7.1	Conclusions	89
7.2	Future Research	91

References

93

List of Figures

1.1	Relationship between chapters	12
2.1	High level Web log mining process	17
3.1	An example of click-through bipartite.	40
3.2	MAP comparisons.	45
4.1	Sequence similarity between two models.	60
4.2	Impact of the parameter α	61
4.3	Impact of the parameter β	61
5.1	Comparison between two models.	75
5.2	The comparison of precision on four classes.	76
5.3	Average precision between two models.	76
6.1	Minimum support and number of the transactions.	87
6.2	The comparison of precision on four classes.	88
6.3	Comparison between two models.	88

List of Tables

3.1	Samples of search engine click-through data	39
3.2	Examples of QCQS query suggestion results	43
3.3	Accuracy comparisons	44
4.1	User profile	53
4.2	Information format of click-through data	57
4.3	Precision comparisons	62
6.1	Web access log	81
6.2	Format of user profile	84

Abstract

With the rapid development of the Internet and WWW, it is more and more important for people to access quality web information. Thus the problem of enabling users to quickly and accurately find information has become an urgent issue. As one of the basic ways to solve this problem, personalized information services have been focusing on fulfilling the personalized information requirements of different users based on their actual demands, preference characteristics, behaviour patterns, etc. This thesis focuses on enhancing web log based recommendation by personalized retrieval, and its main works and innovations include:

- For personalized retrieval, the thesis proposes two models to improve user experience and optimize search performance. The first is a query suggestion model based on query semantics and click-through data. This model calculates the subject relevance between queries, and then combines the semantic information and the relevance of the query-click matrix model as this can effectively eliminate the ambiguity and input errors of reminder queries. The second is a collaborative filtering retrieval model based on local and global features. By the integration of the local and global characteristics of the accessed information, this model overcomes the limitations of a single feature, and increases the degree of application of the retrieval model.
- For recommendation by personalized retrieval, we propose two recommendation models based on the web log. The first is based on the user's atomic

retrieval transaction sequence and the browse characteristics. This model decomposes search transactions, and calculates the user's degree of interest on the search term, which allows users to query information more clearly. Further, it incorporates the user feedback on the search results evaluation value, which overcomes the shortcomings of the model based on content filtering. The second model is based on user interests association findings, which can be used to: find the relationship between resources accessed by users, extract the associations of user interests, and address the problem of user interests isolation.