
Privacy aware social information retrieval and spam filtering using folksonomies

Dissertation zur Erlangung des
naturwissenschaftlichen Doktorgrades
der Bayerischen Julius–Maximilians–Universität Würzburg

vorgelegt von

Beate Navarro Bullock

aus
Nürnberg

Würzburg, 2015

Contents

1	Introduction	1
1.1	Motivation	1
1.2	Problem Statement	5
1.2.1	Social Search	5
1.2.2	Spam Detection	5
1.2.3	Data Privacy	6
1.3	Thesis Outline	6
I	Foundations	9
2	Collaborative Tagging	11
2.1	Characterizing Collaborative Tagging Systems	11
2.1.1	Indexing Documents	11
2.1.2	System Properties	12
2.1.3	Tagging Properties	14
2.2	Formal Model	15
2.3	Tagging Dynamics	18
2.4	Collaborative Tagging and the Semantic Web	20
2.5	Mining Tagging Data	21
2.5.1	Ranking	22
2.5.2	Recommender Systems	24
2.5.3	Community Detection	27
2.6	Example Systems	28
3	Social Information Retrieval	31
3.1	Characterizing Social Information Retrieval	31
3.2	Exploiting Clickdata	33
3.2.1	Query Log Basics	33
3.2.2	Search Engine Evaluation with Query Logs	33
3.2.3	Learning-to-Rank	38
3.2.4	Clickdata as a tripartite Network: Logsonomies	40
3.3	Exploiting Tagging Data	42
3.3.1	Exploiting Bookmarks	42
3.3.2	Exploiting Tags	43

4	Spam Detection	47
4.1	Definition	47
4.2	General Spam Detection Approaches	48
4.2.1	Heuristic Approaches	48
4.2.2	Machine Learning Approaches	49
4.2.3	Evaluation Measures for Spam Filtering	54
4.3	Characterizing Social Spam	56
4.3.1	Spam in Social Bookmarking Systems	56
4.3.2	Spam Detection in other Social Media Applications	59
5	Data Privacy	63
5.1	Basic Concepts	63
5.1.1	Privacy as the Right to Informational Self-Determination	63
5.1.2	Guidelines and Definitions	64
5.1.3	Privacy Principles	65
5.2	Legal Situation in Germany	66
5.2.1	German Federal Protection Act	66
5.2.2	German Federal Telemedia Act	67
5.3	Data Privacy in the Social Web	68
II	Methods	71
6	Social Information Retrieval in Folksonomies and Search Engines	73
6.1	Introduction	73
6.2	Datasets	74
6.2.1	Overview of Datasets	74
6.2.2	Construction of Folk- and Logsonomy Datasets	76
6.2.3	Construction of User Feedback Datasets	77
6.3	Comparison of Searching and Tagging	77
6.3.1	Analysis of Search and Tagging Behaviour	77
6.3.2	Analysis of Search and Tagging System Content	82
6.3.3	Discussion	86
6.4	Properties of Logsonomies	86
6.4.1	Degree distribution	86
6.4.2	Structural Properties	89
6.4.3	Semantic Properties	91
6.4.4	Discussion	95
6.5	Exploiting User Feedback	96
6.5.1	Implicit Feedback from Tagging Data for Learning-to-Rank	97
6.5.2	Mapping click and tagging data to rankings	98
6.5.3	Experimental Setup	99
6.5.4	Discussion	102
6.6	Summary	103

7	Spam Detection in Social Tagging Systems	105
7.1	Introduction	105
7.2	Datasets	106
7.2.1	Dataset Creation	106
7.2.2	Dataset Descriptions	106
7.3	Feature Engineering	109
7.3.1	Feature Description	109
7.3.2	Experimental Setup	113
7.3.3	Results	115
7.3.4	Discussion	117
7.4	ECML/PKDD Discovery Challenge 2008	119
7.4.1	Task Description	119
7.4.2	Methods	120
7.4.3	Results	121
7.4.4	Discussion	121
7.5	Frequent Patterns	122
7.5.1	Quality Functions for Discovering Frequent Spam Patterns	123
7.5.2	Experimental Setup	124
7.5.3	Results	125
7.5.4	Discussion	129
7.6	Summary	130
8	Data Privacy in Social Bookmarking Systems	133
8.1	Introduction	133
8.2	Legal Analysis of Spam Detection	134
8.3	Privacy Aware Spam Experiments	135
8.3.1	Experimental Setup	135
8.3.2	Evaluation	137
8.3.3	Results and Discussion	139
8.4	Summary	140
III	Applications	143
9	BibSonomy Spam Framework	145
9.1	BibSonomy Spam Statistics	145
9.2	Framework Processes and Architecture	146
9.2.1	Generation of the Training Model	148
9.2.2	Classifying New Instances	149
9.3	Implementation Details	150
9.4	Framework Interface	151
9.5	Summary	152
10	Case Study of Data Privacy in BibSonomy	153
10.1	Introduction	153
10.2	Data Privacy Analysis	154
10.2.1	Registration	154

10.2.2	Spam Detection	155
10.2.3	Storing Posts	156
10.2.4	Storing and Processing Publication Metadata	157
10.2.5	Search in BibSonomy	158
10.2.6	Forwarding Data to a Third Party	159
10.2.7	Membership Termination	160
10.2.8	BibSonomy as a Research Project	161
10.3	Discussion	162
11	Conclusion and Outlook	163
11.1	Social Search	163
11.2	Spam Detection	165
11.3	Data Privacy	167
A	Appendix	203

List of Figures

2.1	Elements of the folksonomy	17
2.2	Frequency distributions of Li et al. [2008] and Wetzker et al. [2008] . . .	19
3.1	Example of two weight vectors generated by RankingSVM	41
4.1	Hyperplane which separates positive and negative examples in a multidimensional space	53
4.2	The ROC space and its interpretation	55
6.1	Distribution of items in Delicious and MSN on a log-log scale	79
6.2	Time series of two highly correlated items, "vista" and "iran"	81
6.3	Degree distribution of tags/query words/queries	87
6.4	Degree distribution of resource nodes	88
6.5	Degree distribution of user nodes	89
6.6	Average semantic distance, measured in WordNet, from the original tag to the most closely related one	95
7.1	Histogram of the number of digits in the username and e-mail address . .	110
7.2	ROC curves of the frequency and tf-idf tag features	114
7.3	ROC curves of the different classifiers considering all features	116
7.4	ROC curves of the different feature groups	117
7.5	ROC curves of the two semantic feature groups	118
7.6	AUC values of different submissions	121
9.1	Newly registered spammers and non-spammers having at least one post tracked over time	146
9.2	The three actors (user, framework and administrator) of the spam classification process	147
9.3	Basic components of the spam framework	149
9.4	Administrator interface used to flag BibSonomy users as spammers or legitimate users	151