Contents

1 Introduction **1**

1.1 Problem Setting 4
1.2 Research Objectives 5
1.3 Outline of the Thesis 6

2 Fundamentals **9**

2.1 Research Context 9
2.1.1 Knowledge Discovery and Data Mining 9
2.1.2 Information Retrieval 11
2.2 Structuring: Classification and Clustering 15
2.3 Hierarchies 18
2.3.1 Types of Hierarchies 18
2.3.2 Hierarchical Labeling and Search 20
2.3.3 Hierarchies as Considered in this Thesis 21
2.4 Personalization 22
2.4.1 User Modeling, Adaptivity, and Personalization 22
2.4.2 Personalization in this Thesis 23
2.5 Mining Text Collections 25
2.5.1 Document Preprocessing 25
2.5.2 Feature Selection 27
2.5.3 Document Representation 29
2.5.4 Document Similarity and Distance Measures 31
2.6 Discussion and Conclusion 33

3 Hierarchical Classification **35**

3.1 Objectives 36
3.2 Related Work in Hierarchical Classification 38
3.2.1 Types of Approaches 38
3.2.2 A Review of Specific Approaches 39
3.3 Incomplete Class Structures 41

Bibliografische Informationen
http://d-nb.info/997098635

digitalisiert durch DEUTSCHE NATIONAL BIBLIOTHEK

3.3.1 Rejection . . . 41
3.3.2 Using New Vocabulary . . . 43
3.4 User Centric Hierarchical Classification . . . 44
3.4.1 Prediction Utilities . . . 45
3.4.2 CUClass: A Classification Uncertainty Approach . . . 47
3.4.3 HUClass: A Decision-theoretic Approach . . . 50
3.5 User Centric Performance Evaluation . . . 54
3.5.1 Standard Classification Performance Measures . . . 55
3.5.2 User Centric Performance Measures . . . 56
3.6 Experiments . . . 57
3.6.1 Experimental Setup . . . 58
3.6.2 Naïve Bayes . . . 58
3.6.3 Support Vector Machines (SVMs) . . . 60
3.6.4 An Experimental Comparison of Naïve Bayes Probability Estimates 61
3.6.5 Tuning the Hierarchical Classifiers . . . 61
3.6.6 The Classification Performance of HUClass and CUClass . . . 63
3.6.7 The Runtime Performance of HUClass and CUClass . . . 66
3.6.8 Learning with Very Small Training Sets . . . 68
3.6.9 Incomplete Class Structures . . . 69
3.7 Discussion and Conclusion . . . 76

4 Semi-supervised Hierarchical Clustering 79
4.1 Objectives . . . 80
4.2 Hierarchical Clustering . . . 82
4.3 Constraining Clustering Algorithms . . . 85
4.4 Related Work . . . 90
4.5 Instance-based Constraints in Hierarchical Clustering . . . 94
4.5.1 Integrating MLB Constraints . . . 94
4.5.2 An Efficient Data Structure . . . 96
4.6 Metric Learning for Hierarchical Clustering . . . 98
4.6.1 Weight Learning . . . 100
4.6.2 Incremental Constraints . . . 101
4.6.3 Unevenly Distributed Constraints . . . 102
4.6.4 Exponentially Increasing Constraint Sets . . . 103
4.6.5 Contradicting Constraints . . . 103
4.6.6 Combining mHAC with iHAC . . . 107
4.7 Metric Learning during Hierarchical Clustering . . . 107
4.7.1 The Approach . . . 107

4.7.2 Comparing biHAC with mHAC . . . 110
4.8 Hierarchical Clustering Evaluation . . . 111
4.8.1 Existing Measures . . . 111
4.8.2 Hierarchically Applied F Measure . . . 112
4.8.3 H-Correlation: A New Measure for Hierarchy Comparison . . . 113
4.9 Experiments . . . 116
4.9.1 Experimental Setup . . . 116
4.9.2 Clustering Performance with Evenly Distributed Constraints . . . 118
4.9.3 Clustering Performance with Unevenly Distributed Constraints . . . 120
4.9.4 Runtime Performance . . . 122
4.10 Discussion and Conclusion . . . 125

5 Implementation and Application **129**
5.1 Objectives . . . 129
5.2 CARSA . . . 129
5.2.1 The Search Engine Environment . . . 130
5.2.2 The Evaluation Framework . . . 133
5.3 Discussion of an Application Scenario . . . 134
5.4 Conclusion . . . 138

6 Conclusion **139**
6.1 Summary . . . 139
6.2 Contributions . . . 140
6.3 Limitations and Future Work . . . 142

A Evaluation Datasets **145**
A.1 Banksearch Dataset . . . 145
A.2 Reuters Dataset . . . 145
A.3 Open Directory Dataset . . . 147
A.4 Classification Scenarios with Incomplete Class Structures . . . 148

B Document Preprocessing **149**

C Detailed Clustering Results **151**