Table of Contents

Notation vii

1 Introduction 1

1.1 Thesis Contributions 4

1.2 Thesis Overview 7

2 Document Representation and Retrieval 9

2.1 The Information Retrieval Process 10

2.2 Document Representation 11

2.2.1 Term Vector Models 12

2.2.2 Document Similarity 16

2.2.3 Index Term Set Construction Methods 20

2.3 Techniques for Information Need Satisfaction 23

3 Information Need and Categorizing Search 27

3.1 Supervised vs. Unsupervised Categorization 29

3.2 The Cluster Hypothesis 30

3.3 Clustering Documents 31

3.3.1 Challenges 32

3.3.2 Clustering Algorithms 34

3.4 On the Validity of Document Clusterings 37

3.4.1 External Cluster Validity Measures 38

3.4.2 Internal Cluster Validity Measures 40

3.4.3 Statistical Hypothesis Testing 46

3.4.4 Experimental Evaluation 48

3.4.5 Concluding Remarks 53

3.5 The Suffix Tree Document Representation 54

3.5.1 Suffix Trees 55

3.5.2 A Closer Look to the Suffix Tree Document Model 55
3.5.3 Experimental Evaluation 60
3.5.4 Concluding Remarks 61
3.6 Topic Identification 62
3.6.1 Formal Framework 62
3.6.2 Related Work 64
3.6.3 The WCC Algorithm for Topic Identification 66
3.6.4 Experimental Evaluation 66
3.6.5 Concluding Remarks 70
3.7 Genre Classification 71
3.7.1 What Does Genre Mean? 71
3.7.2 What Does Genre Mean in the WWW? 72
3.7.3 Existing Work 72
3.7.4 User Study and Genre Selection 74
3.7.5 Features for Genre Classification 77
3.7.6 Experimental Evaluation 83
3.7.7 Concluding Remarks 85

4 TIRA: A Software Architecture for Personal IR 87
4.1 From IR Theory to IR Software 88
4.2 Specification of IR Processes 89
4.2.1 Petri Nets 91
4.2.2 Data Flow Graphsand Control/Data Flow Graphs 93
4.2.3 UML Activity Diagrams 94
4.3 Operationalizing IR Processes with TIRA 96
4.3.1 From PIM to PSM 97
4.3.2 The TIRA Middleware Platform 98
4.3.3 TIRA at Work 99
4.4 Concluding Remarks 100

5 Conclusion and Outlook 101