Contents

Abstract **5**

Thesis Details **7**

Acknowledgements **9**

List of Abbreviations **19**

1 Introduction **23**
1.1 Scientific Goal 24
1.2 Challenges 24
1.3 Main contribution 28
1.3.1 Semantic Integration 28
1.3.2 Data Reduction 30
1.3.3 Cross-asset Data Mining 30
1.3.4 Performance Profiling 31
1.3.5 Performance Quantification 31
1.3.6 Our Framework 32
1.4 Validation Domain 32
1.5 Organization of the Thesis 35

2 Literature Review **37**
2.1 Data Mining 37
2.2 Data Integration 40
2.3 Data Reduction 42
2.3.1 Data Reduction in Process Mining 42
2.3.2 Data Reduction in Outlier Detection 43
2.3.3 Data Reduction in Web Log Cleaning 43
2.3.4 Data Reduction in State Index Pruning 44
2.3.5 Data Reduction in Biological Domain 45

2.4 Pattern Mining . . . 45
2.4.1 Sequential Pattern Mining . . . 45
2.4.2 Multilevel Sequential Pattern Mining . . . 46
2.5 Clustering . . . 48
2.6 Classification . . . 51
2.7 Conclusion . . . 53

3 Integration of industrial logs data 55
3.1 Problem Statement and Approach . . . 55
3.1.1 Integration Challenge . . . 56
3.1.2 Documentation Challenge . . . 56
3.1.3 Our Approach . . . 57
3.1.4 Scientific Contribution . . . 58
3.2 Methodology . . . 58
3.2.1 Ontologies . . . 60
3.2.2 Mapping Rules . . . 60
3.2.3 Instantiation . . . 61
3.2.4 Semantic Integration Framework Instantiation . . . 68
3.3 Validation . . . 70
3.4 Discussion and Further Research . . . 72

4 Data Reduction 75
4.1 Problem Statement and Approach . . . 75
4.1.1 Our Approach . . . 77
4.1.2 Scientific Contribution . . . 77
4.2 Methodology . . . 78
4.3 Validation . . . 82
4.3.1 Baseline . . . 83
4.3.2 Experimental Design . . . 86
4.3.3 Evaluation Results . . . 88
4.4 Discussion and Further Research . . . 92

5 Data Exploration Through Multilevel Sequential Pattern Mining 95
5.1 Problem Statement and Approach . . . 95
5.1.1 Data Model Heterogeneity . . . 96
5.1.2 Semantic Heterogeneity . . . 97
5.1.3 Noise . . . 97
5.1.4 Our Approach . . . 97
5.1.5 Scientific Contribution . . . 98
5.2 Methodology . . . 99
5.3 Validation . . . 101
5.3.1 Data Characteristics . . . 101

5.3.2 PMSM vs. APriori . . . 101
5.3.3 PMSM Potential on Heterogeneous Data . . . 103
5.3.4 Data Reduction Impact on Pattern Mining Computation . . . 104
5.4 Discussion and Further Research . . . 107

6 Performance Evaluation of Industrial Assets 109
6.1 Problem Statement and Approach . . . 109
6.1.1 Our Approach . . . 111
6.1.2 Scientific Contribution . . . 111
6.2 Methodology . . . 112
6.3 Validation . . . 115
6.3.1 Evaluation of Event Profiles Representativeness . . . 115
6.3.2 Evaluation of Performance Profile Representativeness . . . 117
6.3.3 Applicability of Performance Profile . . . 122
6.4 Discussion and Further Research . . . 124

7 Performance Labelling Based on Numerically Encoded Event Logs 127
7.1 Problem Statement and Approach . . . 127
7.1.1 Our Approach . . . 128
7.1.2 Scientific Contribution . . . 129
7.2 Methodology . . . 129
7.2.1 Relevancy Score Vectors Annotation . . . 130
7.2.2 Classifying the Relevancy Score Vectors . . . 131
7.2.3 Labelling new operational cycles . . . 131
7.3 Validation . . . 132
7.3.1 Evaluation of the Classifiers . . . 132
7.3.2 Evaluation Against Random Indexing Encoding . . . 134
7.3.3 Evaluation of Performance Classification . . . 135
7.4 Discussion . . . 138

8 Conclusion and Future Research Directions 141
8.1 Further Research . . . 143

Bibliography 145

Appendices 155