

# Table of Contents

Chapter 1: Introduction .....	1
Chapter 2: Background .....	7
2.1 Data Clustering .....	7
2.1.1 Introduction .....	7
2.1.1.1 Practical Applications .....	8
2.1.1.2 Data Clustering is an Ill-posed Task .....	9
2.1.2 The Structure of Data Clustering Output .....	9
2.1.2.1 How Many Clusters .....	9
2.1.2.2 Assignment Probabilities .....	10
2.1.3 Data Representation .....	10
2.1.3.1 Pairwise Representation .....	10
2.1.3.2 Feature-based Representation .....	10
2.1.3.3 Proximity Measures .....	11
2.1.3.4 Re-representation and Preprocessing .....	13
2.1.4 Algorithmic Framework for Data Clustering .....	13
2.1.4.1 Incremental Search .....	14
2.1.4.2 Cost-Based Search .....	15
2.1.4.3 Axiomatic Approach to Data Clustering .....	15
2.1.4.4 Prototypical Representatives of Clusters .....	16
2.1.4.5 Stochasticity .....	17
2.1.4.6 Probabilistic Clustering .....	18
2.1.5 Variations on Data-Clustering .....	20
2.1.5.1 Data Clustering and other Unsupervised Tasks .....	20
2.1.5.2 Methods that Extend Basic Data Clustering .....	21
2.1.5.3 Data Clustering with Constraints .....	22
2.2 Computational Models of Analogy .....	23
2.2.1 The Structure Mapping Theory .....	23
2.2.1.1 Data Representation .....	24
2.2.1.2 Principles and Algorithmic Framework .....	24

2.2.2 The Copycat Project.....	25
2.2.2.1 System Overall Description.....	25
2.2.2.2 Further Discussion in View of Methods Reviewed Previously.....	27
Chapter 3: Setting and Evaluation .....	29
3.1 Problem Setting .....	29
3.2 A Real-world Example .....	31
3.3 Evaluation.....	33
3.3.1 Cluster Purity .....	34
3.3.2 Jaccard coefficient .....	34
3.3.2.1 Probabilistic Extension for Jaccard coefficient .....	36
3.3.2.2 Adapting Jaccard coefficient for the Cross Partition Setting.....	37
Chapter 4: Coupled Clustering.....	39
4.1 Computational Background .....	39
4.1.1 Cost-based Pairwise Clustering .....	39
4.1.2 Feature-based Similarity Measures.....	42
4.2 Algorithmic Framework for Coupled Clustering.....	43
4.2.1 Directing Clustering through Between-subset Similarities.....	43
4.2.2 Three Alternative Coupled Clustering Cost Functions .....	44
4.2.3 Properties of the Coupled Clustering Cost Functions.....	46
4.2.4 Optimization Method.....	48
4.3 Experiments with Synthetic Data .....	49
4.4 Identifying Corresponding Topics in Textual Corpora.....	52
4.4.1 Conflict Keyword Clustering Based on Pre-given Similarities .....	53
4.4.2 Religion Keyword Clustering .....	55
4.4.2.1 The Data .....	56
4.4.2.2 Qualitative Overview of the Result .....	57
4.4.2.3 Expert Data Used for Evaluation .....	58
4.4.2.4 Examples of Expert Data versus Coupled Clustering Output.....	60
4.4.2.5 Quantification of the Overlap with the Expert Data .....	63
4.4.2.6 Agreement between the Experts .....	65
4.5 Discussion.....	66

Chapter 5: Cross-partition Clustering .....	69
5.1 Cross Partition versus Coupled Clustering .....	69
5.2 Background: Information Theoretic Approaches .....	70
5.2.1 The Information Distortion Method.....	70
5.2.1.1 Input and Output.....	70
5.2.1.2 Underlying Principles and Formulation.....	71
5.2.1.3 The ID Algorithm .....	73
5.2.1.4 Controlling the Number of Clusters by Modifying the Value of $\beta$ .....	75
5.2.2 The Information Bottleneck Method .....	76
5.2.2.1 The IB Method and Information Theory .....	77
5.2.3 Information Bottleneck with Side Information.....	78
5.3 The Cross-partition Method.....	80
5.3.1 The Cross-partition Data Clustering Task .....	81
5.3.1.1 Input: The Pre-partitioning Variable.....	81
5.3.1.2 Output: Re-association of Features and Clusters .....	81
5.3.2 Underlying Principles Characterizing the Solution .....	82
5.3.2.1 Assignments of Elements to Clusters .....	83
5.3.2.2 $W$ -projected Centroids .....	83
5.3.2.3 Feature-cluster Re-association.....	84
5.3.2.4 Centroids that Cut Across the Pre-partition.....	85
5.3.3 The CP Algorithm.....	85
5.3.3.1 Further Observations .....	87
5.3.3.2 The Parameters $\beta$ and $\eta$ .....	88
5.3.4 CP Algorithmic Variations Inspired by the IB Method.....	90
5.4 Experimental Work.....	91
5.4.1 Experiments with Synthetic Data.....	91
5.4.1.1 Setting.....	92
5.4.1.2 Results .....	93
5.4.1.3 Oscillatory Endless Loops .....	95
5.4.2 Application to Religion Data .....	97
5.4.2.1 Results .....	97
5.4.2.2 Quantification of the Overlap with the Expert Data .....	100
5.4.2.3 Agreement between the Experts .....	105

5.5 Discussion.....	106
Chapter 6: Discussion and Further Directions .....	109
Appendix A: Religion Data .....	117
A.1 The Corpora .....	117
A.2 The Features.....	118
A.3 The Clustered Keyword Sets .....	120
Appendix B: Examples of Religion Coupled Clustering.....	123
Appendix C: The Expert Data.....	129
C.1 Instructions for Participants .....	129
C.2 Religion-Related Term Classes Contributed by Experts.....	131
C.2.1 The Data Contributed by Expert I.....	131
C.2.2 The Data Contributed by Expert II.....	132
C.2.3 The Data Contributed by Expert III .....	134
Appendix D: Proofs for Chapter 5 .....	135
Appendix E: Examples of Religion Cross Partition Clustering.....	141
E.1: Two Clusters .....	141
E.2: Seven Clusters.....	145
References.....	151