

Contents

1	Introduction	1
1.1	Motivation	1
1.2	Thesis and goals of the dissertation	2
1.3	Organization of the dissertation	3
2	Biclustering	4
2.1	Background	4
2.1.1	Definition	4
2.1.2	Classification of biclusters	5
2.2	Biclustering proximity measures	8
2.2.1	Distance-based measures	9
2.2.2	Qualitative measures	9
2.2.3	Non-correlation-based measures	10
2.2.4	Correlation-based measures	10
2.3	Biclustering algorithms	13
2.3.1	Cheng-Church (CC)	15
2.3.2	Plaid Model	15
2.3.3	Order-Preserving Submatrix (OPSM)	16
2.3.4	ISA	17
2.3.5	xMotifs	18
2.3.6	Bimax	18
2.3.7	Bayesian Biclustering (BBC)	20
2.3.8	Correlated Pattern Biclustering (CPB)	20
2.3.9	Qualitative biclustering (QUBIC)	21
2.3.10	Factor analysis for bicluster acquisition (FABIA)	23
2.3.11	Other algorithms	24
2.3.12	Ensemble methods	24
2.4	Biclustering validation	25
2.4.1	Internal indices	25
2.4.2	External indices	26
2.4.3	Relative indices	27
2.4.4	Visualization: methods & tools	28
2.5	Summary	29
3	Common biclustering datasets	30
3.1	Synthetic datasets	30
3.2	Biological datasets - microarrays	31
3.2.1	Background	31
3.2.2	Microarray technology	33
3.2.3	Spotted (cDNA) microarrays	34

3.2.4	Oligonucleotide microarrays	34
3.2.5	Summary	36
3.3	Microarray Data Analysis	36
3.3.1	Microarray data availability	36
3.3.2	Preprocessing microarray data	37
3.4	Other datasets	41
3.4.1	Social network datasets	41
3.5	Summary	41
4	Algorithms	42
4.1	Propagation-Based Biclustering Algorithm (PBBA)	43
4.1.1	Definitions	43
4.1.2	Algorithm	46
4.1.3	Complexity	48
4.1.4	Summary	49
4.2	Modifications and adaptations of PBBA	49
4.2.1	Qualitative Propagation Biclustering (QPB)	49
4.2.2	Algorithm for Business Continuity Plan Pitfalls Prevention (ABCPPP)	50
4.3	MiniMax with Pearson Correlation (MMPC)	53
4.3.1	The algorithm	54
4.3.2	Complexity	54
4.3.3	Summary	54
4.4	Modifications of MMPC	56
4.4.1	Maximal Pearson Correlation (MPC)	56
4.4.2	MiniMax with Spearman Correlation (MMSC)	56
4.5	Summary	56
5	Environments and implementation	57
5.1	Computer architecture and organization	57
5.1.1	Background	58
5.1.2	Digital logic	59
5.1.3	Microarchitecture	60
5.1.4	CPU architecture	62
5.1.5	Memory architecture	65
5.1.6	GPU architecture	67
5.2	Tools and programming patterns	72
5.2.1	Libraries	72
5.2.2	Sequential patterns	74
5.2.3	Parallel patterns	76
5.3	Efficient Implementation	79
5.3.1	Propagation-Based Biclustering Algorithm (PBBA)	79
5.3.2	Min-Max with Pearson Correlation (MMPC)	80
5.3.3	Cheng-Church algorithm	84

5.3.4	Bimax algorithm	85
5.3.5	Other scripts	85
5.4	Summary	86
6	Results	87
6.1	Background	87
6.1.1	Amdahl's Law	88
6.1.2	Gustafson's Law	89
6.1.3	Criticism and equivalence.	89
6.2	Comparison methodology	90
6.3	Propagation-Based Biclustering Algorithm (PBBA)	92
6.3.1	Verification test	92
6.3.2	Biological relevance	94
6.3.3	Summary	97
6.4	MiniMax with Pearson Correlation (MMPC)	98
6.4.1	Results on different architectures	98
6.4.2	Conclusions	100
6.4.3	Summary	101
6.5	Comparison experiment	102
6.5.1	Details of the approach	102
6.5.2	Biclustering algorithms parameters	103
6.5.3	Biological validation	105
6.5.4	The results	106
6.5.5	Conclusions	111
6.5.6	Criticism	113
6.6	Novel areas of biclustering application	114
6.6.1	Gowalla	114
6.6.2	Twitter	116
6.6.3	Summary	121
6.7	Integrations	122
6.8	Summary	122
7	Concluding remarks	123
7.1	Contribution	123
7.2	Discussion	124
7.3	Future work	125
	Bibliography	126
A	Appendix: Intel MIC Architecture	144
B	Appendix: Detailed results	148
C	Appendix: Properties of GPUs	162