

Contents

Acknowledgments	ii
Abstract	iii
List of Figures	viii
List of Tables	x
1 Introduction to Data Mining	1
1.1 Data Mining and Knowledge Discovery	3
1.1.1 Data mining - On What Kind of Data?	6
1.2 Data Mining Tasks	7
1.3 Requirements and Challenges of Data Mining	12
1.4 Related Work and Thesis Contributions	13
1.5 Thesis Outline	15
2 Problem Formulation and Data Representation	16
2.1 Problem Setting and Preliminaries	16
2.2 Search-Space	19
2.3 Data Representation	21
3 Horizontal Approach	24
3.1 Horizontal Algorithms for Mining Frequent Itemsets	24

3.1.1	Apriori Algorithm	25
3.1.2	FP-growth Algorithm	28
3.2	Horizontal Algorithms for Mining Sequential Patterns	34
3.2.1	GSP Algorithm	35
3.2.2	PrefixSpan Algorithm	37
4	Vertical Approach	41
4.1	Vertical Tidset-Based Approach	41
4.1.1	Lattice Decomposition: Prefix-Based Classes	43
4.1.2	Patterns Generation	44
4.1.3	Enumerating Frequent Patterns of a Class	46
4.1.4	Frequency Counting	48
4.1.5	Spade Algorithm	51
4.1.6	Eclat Algorithm	51
4.1.7	Diffset Data Representation	53
4.1.8	dEclat Algorithm	55
4.2	Vertical Bitmap-Based Approach	57
5	New DATA REPRESENTATIONS	60
5.1	Primeset Data Representation	61
5.1.1	Preliminaries Concepts	61
5.1.2	Introducing Primeset	63
5.1.3	Joining Primesets	65
5.1.4	PrimeEclat Algorithm	67
5.1.5	Optimizations	69
5.1.6	Introducing PS-List	72
5.1.7	Joining PS-Lists	74

5.1.8	PrimeSpade Algorithm	77
5.2	Diffseq Data Representation	80
5.2.1	Introducing Diffseq	81
5.2.2	Joining Diffseqs	81
5.2.3	dSpade Algorithm	85
6	Experimental Evaluation	88
6.1	Datasets	89
6.2	Effects of Optimizations	90
6.3	Experimental Results	90
6.3.1	PrimeEclat Performance	91
6.3.2	PrimeSpade and dSpade Performance	92
7	Conclusions and Future Work	106
7.1	Conclusions	106
7.2	Future Work	107
	Bibliography	108
A	Euclidean Algorithm	113