

Contents

1	Introduction	1
1.1	Motivation	1
1.2	Problem Definition and Main Objectives	2
1.3	Main Contributions	3
1.4	Conclusions	5
	References	5
2	Related Work and Concepts	7
2.1	Processing Complex Data	7
2.2	Knowledge Discovery in Traditional Data	9
2.3	Clustering Complex Data	10
2.4	Labeling Complex Data	14
2.5	MapReduce	16
2.6	Conclusions	17
	References	17
3	Clustering Methods for Moderate-to-High Dimensionality Data	21
3.1	Brief Survey	21
3.2	CLIQUE	22
3.3	LAC	25
3.4	CURLER	26
3.5	Conclusions	29
	References	31
4	Halite	33
4.1	Introduction	33
4.2	General Proposal	35
4.3	Presented Method: Basics	37
4.3.1	Building the Counting-Tree	38
4.3.2	Finding β -Clusters	41
4.3.3	Building the Correlation Clusters	47

4.4	Presented Method: The Algorithm Halite	47
4.5	Presented Method: Soft Clustering.	49
4.6	Implementation Discussion	52
4.7	Experimental Results	53
4.7.1	Comparing Hard Clustering Approaches	53
4.7.2	Comparing Soft Clustering Approaches	61
4.7.3	Scalability	63
4.7.4	Sensitivity Analysis	63
4.8	Discussion	64
4.9	Conclusions	66
	References	66
5	BoW	69
5.1	Introduction	69
5.2	Main Ideas of BoW: Reducing Bottlenecks	71
5.2.1	Parallel Clustering: ParC	72
5.2.2	Sample and Ignore: SnI	73
5.3	Cost-Based Optimization of BoW	76
5.4	Finishing Touches: Data Partitioning and Cluster Stitching.	80
5.4.1	Random-Based Data Partition.	80
5.4.2	Location-Based Data Partition	80
5.4.3	File-Based Data Partition	82
5.5	Experimental Results	83
5.5.1	Comparing the Data Partitioning Strategies	85
5.5.2	Quality of Results	86
5.5.3	Scale-Up Results.	87
5.5.4	Accuracy of the Cost Equations	89
5.6	Conclusions	91
	References	91
6	QMAS	93
6.1	Introduction	93
6.2	Presented Method	95
6.2.1	Mining and Attention Routing	95
6.2.2	Low-Labor Labeling (LLL)	99
6.3	Experimental Results	101
6.3.1	Results on the Initial Example	102
6.3.2	Speed	103
6.3.3	Quality and Non-labor Intensive	104
6.3.4	Functionality	104
6.3.5	Experiments on the <i>SATLARGE</i> Dataset	106
6.4	Conclusions	108
	References	108

7 Conclusion	111
7.1 Main Contributions	112
7.2 Discussion	113
References	115

Data Mining in Large Sets of Complex Data

Ferreira Cordeiro, R.L.; Faloutsos, C.; Traina Júnior, C.

2013, XI, 116 p. 37 illus., 25 illus. in color., Softcover

ISBN: 978-1-4471-4889-0