

Contents

1	Introduction	1
1.1	Introduction	1
1.2	Problem	3
1.3	Hypothesis and Methodology	5
2	Background	6
2.1	Graph Theory	6
2.2	RDF	7
2.2.1	Isomorphism in RDF	9
2.2.2	Entailment and Equivalence in RDF	10
2.2.3	Canonicalisation of RDF Graphs	12
2.3	SPARQL	14
2.3.1	Algebra	15
2.3.2	Semantics	16
2.3.3	Set or Bag Semantics	16
2.3.4	Other Features	17
2.3.5	Other Query Types	19
2.3.6	Query Equivalence and Containment	19
3	Canonicalisation of SPARQL queries	21
3.1	Overview	21
3.2	Supported Queries	22
3.3	r-Graph for Monotone Queries	22
3.3.1	Terms	23
3.3.2	Triple Patterns	23
3.3.3	Joins and Unions	24
3.3.4	Projection	24
3.4	UCQ Normal Form	25
3.4.1	Formalisation of UCQs	26
3.4.2	Projection with Union	28
3.4.3	Redundancy-free UCQs	32
3.4.4	Canonical Variables	35
3.4.5	Canonical Query	36
3.5	Computational Properties	37
3.6	Other Features	37
3.6.1	OPTIONAL	38

3.6.2	FILTER	38
3.6.3	ORDER BY	40
3.6.4	OFFSET and LIMIT	41
3.6.5	Named Graphs	41
3.6.6	GRAPH	42
3.6.7	Canonicalisation of Other Features	42
4	Implementation	44
4.1	Parsing the Query	44
4.2	Normalisation	46
4.3	Canonical Labelling	46
4.4	Minimisation	49
4.5	Generation of Queries	50
4.6	Testing	51
5	Results and Evaluation	53
5.1	Queries Used	53
5.1.1	Real-world Queries	53
5.1.2	Synthetic CQs	54
5.1.3	Synthetic UCQs	56
5.2	Machine Specifications	57
5.3	Results	57
5.3.1	Real-world Queries	57
5.3.2	Synthetic CQs	58
5.3.3	Synthetic UCQs	59
5.4	Discussion	60
6	Conclusion	62
6.1	Summary	62
6.2	Lessons Learned	63
6.3	Future Work	63
	Bibliography	65
	Appendix A	67