

# Contents

<b>1. Introduction</b>	<b>1</b>
1.1. Hypothesis and Objectives . . . . .	3
1.1.1. Hypothesis . . . . .	4
1.1.2. Objectives . . . . .	4
1.2. Methodology . . . . .	5
1.3. Contribution of this Work . . . . .	6
1.4. Structure of This Work . . . . .	7
<b>2. Preliminaries</b>	<b>8</b>
2.1. The Semantic Web . . . . .	8
2.1.1. RDF . . . . .	8
2.1.2. RDFS and OWL . . . . .	9
2.1.3. SPARQL . . . . .	11
2.2. Similarity Search . . . . .	13
2.2.1. Vector Spaces . . . . .	13
2.2.2. Metric Spaces . . . . .	14
2.2.3. Similarity Queries . . . . .	14
2.2.4. Similarity Joins . . . . .	15
<b>3. Related Work</b>	<b>16</b>
3.1. Similarity Join Evaluation . . . . .	16
3.2. Similarity in Databases . . . . .	19
<b>4. An Efficient Algorithm for Approximated Similarity Joins</b>	<b>24</b>
4.1. Preliminary Work . . . . .	24
4.2. The Algorithm . . . . .	25
4.2.1. Complexity Analysis . . . . .	27
4.2.2. Implementation . . . . .	28
4.3. Evaluation . . . . .	29
4.3.1. Experimental Settings . . . . .	29
4.3.2. Average Precision . . . . .	29
4.3.3. Precision Distribution . . . . .	31
4.3.4. Execution Time and Distance Calculations . . . . .	32
4.3.5. Centre Selection Strategies . . . . .	33
4.3.6. Comparison with Other Similarity Join Algorithms . . . . .	35
<b>5. Similarity Joins in SPARQL Queries</b>	<b>37</b>

5.1.	Desiderata . . . . .	37
5.2.	Challenges . . . . .	38
5.2.1.	Size of the Data . . . . .	38
5.2.2.	Dimension-Agnostic Indexing . . . . .	38
5.2.3.	Multi-valued Attributes . . . . .	39
5.2.4.	Incomplete Data . . . . .	40
5.2.5.	Value Distribution . . . . .	40
5.2.6.	Abstract Distances . . . . .	41
5.3.	Syntax . . . . .	41
5.4.	Semantics . . . . .	42
5.5.	Algebraic Properties . . . . .	44
5.6.	Use-Case Queries . . . . .	46
5.7.	Implementation . . . . .	48
5.7.1.	Implementation Details . . . . .	49
5.8.	Query Planning and Optimisation . . . . .	51
5.9.	Evaluation . . . . .	53
5.9.1.	Wikidata: k-nn Self-Similarity Queries . . . . .	54
5.9.2.	Corel Colour Moments: Range Similarity Queries . . . . .	56
<b>6.</b>	<b>Applications</b>	<b>58</b>
6.1.	IMGpedia . . . . .	58
6.1.1.	Browsing the Data . . . . .	59
6.1.2.	Querying the Data . . . . .	61
6.2.	Dynamic Clustering of SPARQL Query Results . . . . .	66
6.2.1.	Syntax and Semantics . . . . .	67
6.2.2.	Implementation . . . . .	68
6.2.3.	Use-Cases . . . . .	70
<b>7.</b>	<b>Conclusions and Future Directions</b>	<b>72</b>
7.1.	Conclusions . . . . .	72
7.2.	Assessment of the Objectives of the Thesis . . . . .	73
7.3.	Future Work . . . . .	74
	<b>Bibliography</b>	<b>76</b>