Extending Knowledge Bases from Human-readable Tables

Kruit, Benjamin Bienze

2021

document version
Publisher's PDF, also known as Version of record

Contents

1 Introduction **11**

1.1 Integrating Tables with Knowledge Bases 14

1.2 Research Questions and Contributions 16

1.2.1 Extracting New Facts 16

1.2.2 Integrating N-ary Tables 17

1.2.3 Creating a New KB from Tables 18

1.2.4 Designing Pipelines 19

1.3 Thesis Outline and Publications 20

2 Background and Related Work **23**

2.1 Tabular Data 24

2.1.1 Data Modeling 24

2.1.2 Web Tables 28

2.1.3 Web Table Extraction 29

2.1.4 Web Table Structures 30

2.2 Knowledge Bases 33

2.2.1 Overview of Knowledge Bases 33

2.2.2 Representing N-ary Information 36

2.2.3 Ontology Matching 38
2.2.4 Knowledge Acquisition 39
2.3 Table Interpretation 41
2.3.1 Common Assumptions 42
2.3.2 Approaches and Systems 44
2.3.3 Constrained Table Interpretation 46
2.4 Data Integration at Scale 48
2.4.1 Schema Matching and Profiling 48
2.4.2 Instance Matching and Entity Linking 51
2.5 Impact and Applications 52
2.5.1 Search, Query Answering and Integration 53
2.5.2 Dataspaces 54
2.6 Summary 55

3 Extracting Novel Facts from Tables 57
3.1 Introduction 58
3.2 Motivation: Measuring Redundancy 59
3.2.1 Experiments 61
3.3 Our Approach 65
3.3.1 Background 65
3.3.2 Intuition 67
3.3.3 Table Interpretation 69
3.3.4 Slot-Filling 75
3.4 Evaluation 76
3.4.1 Table Interpretation 79
3.4.2 Measuring Redundancy 81
3.4.3 Slot-filling 83
3.5 Conclusion and Discussion 84

4 Extracting N-ary Facts from Table Clusters 87
4.1 Introduction 88

4.2 Background 90
4.3 Method 94
4.3.1 Reshaping 95
4.3.2 Clustering 101
4.3.3 Integration 107
4.4 Evaluation 109
4.4.1 Annotations 110
4.4.2 Table Reshaping 111
4.4.3 Table Clustering 112
4.4.4 Table Interpretation and KB Integration 113
4.5 Conclusion 114

5 Building a KB from Tables in Scientific Papers 117
5.1 Introduction 118
5.2 Background 120
5.3 Overview 123
5.4 Task 1: Table Interpretation 126
5.4.1 Training Data Generation 126
5.4.2 Table Header Detection 128
5.4.3 Table Type Detection 128
5.4.4 Column Type Detection 129
5.5 Task 2: Entity Linking 130
5.6 Evaluation 134
5.6.1 Table Interpretation 136
5.6.2 Entity Linking 139
5.7 Conclusion 140

6 A Platform for Web Table Information Extraction 141
6.1 Introduction 142
6.2 Design of Takco 143
6.2.1 Target Users 146

6.3 Usage of Takco 147
6.3.1 Usage as an Analysis Tool 149
6.3.2 Usage as a Modular Platform 151
6.4 Evaluation Results 152
6.5 Conclusion 154

7 Conclusion **157**
7.1 Main Contributions and Future Work 157
7.1.1 Extracting Novel Facts 157
7.1.2 Integrating N-ary Tables 160
7.1.3 Creating a New KB from Tables 162
7.1.4 Designing Pipelines 164
7.2 Final Reflections 166

Bibliography **168**

Summary **191**

Curriculum Vitae **193**