

# Contents

## Part I Overview

<b>1</b>	<b>Introduction</b>	3
1.1	Aims and Challenges of Data Matching	3
1.1.1	Lack of Unique Entity Identifiers and Data Quality	5
1.1.2	Computation Complexity	5
1.1.3	Lack of Training Data Containing the True Match Status	6
1.1.4	Privacy and Confidentiality	6
1.2	Data Integration and Link Analysis	6
1.3	A Short History of Data Matching	9
1.4	Example Application Areas	11
1.4.1	National Census	11
1.4.2	The Health Sector	12
1.4.3	National Security	13
1.4.4	Crime and Fraud Detection and Prevention	14
1.4.5	Business Mailing Lists	15
1.4.6	Bibliographic Databases	17
1.4.7	Online Shopping	18
1.4.8	Social Sciences and Genealogy	19
1.5	Further Reading	20
<b>2</b>	<b>The Data Matching Process</b>	23
2.1	Overview	23
2.1.1	A Small Data Matching Example	23
2.2	Data Pre-Processing	24
2.3	Indexing	27
2.4	Record Pair Comparison	29

2.5	Record Pair Classification . . . . .	32
2.6	Evaluation of Matching Quality and Complexity . . . . .	34
2.7	Further Reading . . . . .	35

## Part II Steps of the Data Matching Process

<b>3</b>	<b>Data Pre-Processing . . . . .</b>	<b>39</b>
3.1	Data Quality Issues Relevant to Data Matching . . . . .	39
3.2	Issues with Names and Other Personal Information . . . . .	42
3.3	Types and Sources of Variations and Errors in Names . . . . .	45
3.4	General Data Cleaning Tasks . . . . .	48
3.5	Data Pre-Processing for Data Matching . . . . .	51
3.5.1	Removing Unwanted Characters and Tokens . . . . .	51
3.5.2	Standardisation and Tokenisation . . . . .	53
3.5.3	Segmentation into Output Fields . . . . .	55
3.5.4	Verification . . . . .	56
3.6	Rule-Based Segmentation Approaches . . . . .	58
3.7	Statistical Segmentation Approaches . . . . .	60
3.7.1	Hidden Markov Model Based Segmentation . . . . .	62
3.8	Practical Considerations and Research Issues . . . . .	65
3.9	Further Reading . . . . .	66
<b>4</b>	<b>Indexing . . . . .</b>	<b>69</b>
4.1	Why Indexing? . . . . .	69
4.2	Defining Blocking Keys . . . . .	70
4.3	(Phonetic) Encoding Functions . . . . .	74
4.3.1	Soundex . . . . .	74
4.3.2	Phonex . . . . .	75
4.3.3	Phonix . . . . .	76
4.3.4	NYSIIS . . . . .	76
4.3.5	Oxford Name Compression Algorithm . . . . .	77
4.3.6	Double-Metaphone . . . . .	78
4.3.7	Fuzzy Soundex . . . . .	78
4.3.8	Other Encoding Functions . . . . .	79
4.4	Standard Blocking . . . . .	80
4.5	Sorted Neighbourhood Approach . . . . .	81
4.6	Q-Gram Based Indexing . . . . .	84
4.7	Suffix-Array Based Indexing . . . . .	86
4.8	Canopy Clustering . . . . .	89
4.9	Mapping Based Indexing . . . . .	92
4.10	A Comparison of Indexing Techniques . . . . .	93
4.11	Other Indexing Techniques . . . . .	94

4.12	Learning Optimal Blocking Keys . . . . .	97
4.13	Practical Considerations and Research Issues . . . . .	98
4.14	Further Reading . . . . .	100
<b>5</b>	<b>Field and Record Comparison . . . . .</b>	<b>101</b>
5.1	Overview and Motivation . . . . .	101
5.2	Exact, Truncate and Encoding Comparison . . . . .	102
5.3	Edit Distance String Comparison . . . . .	103
5.3.1	Smith-Waterman Edit Distance String Comparison . . . . .	105
5.4	$Q$ -gram Based String Comparison. . . . .	106
5.5	Jaro and Winkler String Comparison . . . . .	109
5.6	Monge-Elkan String Comparison . . . . .	111
5.7	Extended Jaccard Comparison . . . . .	112
5.8	SoftTFIDF String Comparison . . . . .	113
5.9	Longest Common Substring Comparison. . . . .	114
5.10	Other Approximate String Comparison Techniques. . . . .	116
5.10.1	Bag Distance . . . . .	116
5.10.2	Compression Distance . . . . .	116
5.10.3	Editex . . . . .	117
5.10.4	Syllable Alignment Distance . . . . .	118
5.11	String Comparison Examples . . . . .	118
5.12	Numerical Comparison . . . . .	121
5.13	Date, Age and Time Comparison . . . . .	122
5.14	Geographical Distance Comparison. . . . .	124
5.15	Comparing Complex Data . . . . .	124
5.16	Record Comparison . . . . .	125
5.17	Practical Considerations and Research Issues. . . . .	126
5.18	Further Reading . . . . .	127
<b>6</b>	<b>Classification . . . . .</b>	<b>129</b>
6.1	Overview. . . . .	129
6.2	Threshold-Based Classification. . . . .	131
6.3	Probabilistic Classification. . . . .	133
6.4	Cost-Based Classification . . . . .	137
6.5	Rule-Based Classification . . . . .	139
6.6	Supervised Classification Methods . . . . .	142
6.7	Active Learning Approaches . . . . .	147
6.8	Managing Transitive Closure . . . . .	149
6.9	Clustering-Based Approaches. . . . .	150
6.10	Collective Classification . . . . .	154
6.11	Matching Restrictions and Group Linking . . . . .	157
6.12	Merging Matches . . . . .	160

6.13	Practical Considerations and Research Issues . . . . .	161
6.14	Further Reading . . . . .	162
<b>7</b>	<b>Evaluation of Matching Quality and Complexity . . . . .</b>	<b>163</b>
7.1	Overview. . . . .	163
7.2	Measuring Matching Quality . . . . .	165
7.3	Measuring Matching Complexity . . . . .	172
7.4	Clerical Review . . . . .	174
7.5	Public Test Data. . . . .	176
7.6	Synthetic Test Data. . . . .	178
7.7	Practical Considerations and Research Issues. . . . .	183
7.8	Further Reading . . . . .	184
 <b>Part III Further Topics</b>		
<b>8</b>	<b>Privacy Aspects of Data Matching. . . . .</b>	<b>187</b>
8.1	Privacy and Confidentiality Challenges for Data Matching . . .	187
8.1.1	Requiring Access to Identifying Information . . . . .	188
8.1.2	Sensitive and Confidential Outcomes from Matched Data. . . . .	189
8.2	Data Matching Scenarios. . . . .	190
8.3	Privacy-Preserving Data Matching Techniques. . . . .	193
8.3.1	Exact Privacy-Preserving Matching Techniques . . . . .	196
8.3.2	Approximate Privacy-Preserving Matching Techniques . . . . .	199
8.3.3	Scalable Privacy-Preserving Matching Techniques. . . . .	203
8.4	Practical Considerations and Research Issues. . . . .	205
8.5	Further Reading . . . . .	207
<b>9</b>	<b>Further Topics and Research Directions . . . . .</b>	<b>209</b>
9.1	Geocode Matching . . . . .	209
9.2	Matching Unstructured and Complex Data . . . . .	211
9.3	Real-time Data Matching. . . . .	213
9.4	Matching Dynamic Databases . . . . .	215
9.5	Parallel and Distributed Data Matching. . . . .	217
9.6	Research Challenges and Directions . . . . .	222
<b>10</b>	<b>Data Matching Systems . . . . .</b>	<b>229</b>
10.1	Commercial Systems and Checklist . . . . .	229
10.2	Research and Open Source Systems . . . . .	231
10.2.1	BigMatch. . . . .	231
10.2.2	D-Dupe . . . . .	232

10.2.3	DuDe . . . . .	232
10.2.4	FEBRL . . . . .	234
10.2.5	FRIL . . . . .	236
10.2.6	Merge ToolBox . . . . .	238
10.2.7	OYSTER . . . . .	239
10.2.8	R RecordLinkage . . . . .	240
10.2.9	SecondString . . . . .	240
10.2.10	SILK . . . . .	240
10.2.11	SimMetrics . . . . .	241
10.2.12	TAILOR . . . . .	241
10.2.13	WHIRL . . . . .	241
<b>Glossary . . . . .</b>		<b>243</b>
<b>References . . . . .</b>		<b>251</b>
<b>Index . . . . .</b>		<b>265</b>

Data Matching

Concepts and Techniques for Record Linkage, Entity  
Resolution, and Duplicate Detection

Christen, P.

2012, XX, 272 p., Hardcover

ISBN: 978-3-642-31163-5