

Inhaltsverzeichnis

Zusammenfassung	9
Vorwort	10
1. Einführung	11
1.1 Motivation und Problemstellung.....	11
1.2 Ziel der Arbeit	20
1.3 Aufbau der Arbeit.....	21
2. Grundlagen	22
2.1 Softwarearchitektur.....	22
2.2 Relationales Datenbankmanagementsystem (RDBMS)	25
2.2.1 Einführung.....	25
2.2.2 ANSI/SPARC 3-Ebenen Modell	28
2.2.3 Referenzarchitektur eines DBMS.....	30
2.2.4 Relationale Datenbanken.....	34
2.2.5 Relationale Anfragesprache.....	37
2.2.5.1 Einführung	37
2.2.5.2 Semantische Integritätsbedingungen	40
2.2.5.3 Prozedurale Sprachelemente	46
2.3 Datenqualität.....	48
2.3.1 Einführung	48
2.3.2 Datenfehlerkategorien.....	49
2.3.3 Definition	51
2.3.4 Managementansätze zur Datenqualität	53
2.3.4.1 TDQM (Wang)	53
2.3.4.2 TQdM (English)	54
2.3.5 Datenqualitätsmerkmale	55

2.3.5.1	Datenqualitätsmerkmale und Metriken	55
2.3.5.2	Klassifikationsansätze	61
3	Aufgaben einer Datenqualitätskomponente	67
3.1	Problematik	67
3.2	Grundlagen.....	68
3.2.1	Codes	68
3.2.1.1	Grundlagen.....	68
3.2.1.2	Patterncodes	69
3.2.1.3	Phonetische Codes	70
3.2.1.4	Prüfzifferncodes.....	75
3.2.1.5	Hashingcodes	79
3.2.1.6	Stemming	81
3.2.1.7	Stoppwörter	83
3.2.1.8	Codelisten.....	84
3.2.2	Proximitätsmaße	85
3.2.2.1	Grundlagen	85
3.2.2.2	Numerische Daten	85
3.2.2.3	Alphanumerische Daten	86
3.2.3	Tokenizer	103
3.2.4	Nachschlagetabellen	105
3.2.5	Reguläre Ausdrücke	107
3.2.6	Segmentierung	108
3.2.7	Klassifizierer	109
3.2.8	Regeln	111
3.2.8.1	Einführung	111
3.2.8.2	Regelspezifikation	112
3.2.8.3	Regelinduktion	115
3.3	Stichprobenentnahme (Sampling)	119

3.4 Verstehen	120
3.4.1 Grundlagen	120
3.4.2 Analysieren	120
3.4.3 Regeln	128
3.5 Verifizieren und Verbessern	130
3.5.1 Standardisieren	130
3.5.2 Anreichern	131
3.5.3 Duplikate	132
3.5.3.1 Einführung	132
3.5.3.2 Reduktion des Suchraums	134
3.5.3.3 Duplikate erkennen	139
3.5.3.4 Duplikate zusammenführen	144
3.5.3.5 Metriken zur Überprüfung	149
3.5.4 Validierung	153
3.6 Steuern	154
3.6.1 Messen von Datenqualitätsmerkmalen	154
3.6.2 Aggregation von Datenqualitätsdimensionen	155
3.6.2.1 Datenqualitätspyramide	155
3.6.2.2 Simple Additive Weighting (SAW)	157
3.6.2.3 Weighted Product	159
3.6.3 Monitoring	159
4 Integration der Datenqualitätsverfahren	160
5 Architektur der Datenqualitätskomponente	164
5.1 Architektur	164
5.2 Klassenbibliothek	166
5.3 Regeln	169
5.4 Duplikate	171

6 Implementation eines Prototyps.....	172
6.1 Allgemein	172
6.2 Basis	174
6.2.1 Allgemein	174
6.2.2 Encoder.....	174
6.2.3 Matching.....	181
6.2.4 Tokenizer	187
6.2.5 Erweiterungen	188
6.3 Verstehen	193
6.3.1 Allgemein	193
6.3.2 Daten- und Schemaeigenschaften.....	193
6.3.3 Primär- und Fremdschlüssel	198
6.3.4 Regelinduktion	200
6.3.5 Erweiterungen	201
6.4 Verbessern	203
6.4.1 Allgemein	203
6.4.2 Regelverwaltung	204
6.4.3 Daten standardisieren, korrigieren, anreichern	210
6.4.4 Duplikaterkennung und -fusion	212
6.4.5 Erweiterungen	220
6.5 Steuern	224
6.5.1 Allgemein	224
6.5.2 Datenqualitätsdimensionen und Regeln	224
7 Schlussbemerkung und Kritik.....	225
Literatur	230