Ein metrikbasierter Ansatz zur Messung der Aktualität von Daten in Informationssystemen

Bernd Heinrich, Mathias Klier, Quirin Görz

Zusammenfassung: Die Verbesserung der Aktualität von Daten in Informationssystemen wird in Wissenschaft und Praxis intensiv diskutiert. In diesem Zuge werden auch geeignete Metriken zur Messung der Aktualität von Daten gefordert. Deshalb wird im Beitrag eine wahrscheinlichkeitstheoretisch fundierte Metrik zur weitgehend automatisierbaren Messung der Aktualität konstruiert, die im Vergleich zu bestehenden Ansätzen eine Kardinalskalierung und Interpretierbarkeit der Metrikergebnisse als Wahrscheinlichkeiten gewährleistet. Damit können die Metrikergebnisse methodisch fundiert in Erwartungswertkalküle von Entscheidungen eingehen. Ferner erlaubt die Metrik eine Konfiguration, um v. a. datenattributspezifische Charakteristika und vorhandene Zusatzdaten bei der Messung zu berücksichtigen. Die Evaluation des Ansatzes erfolgt einerseits anhand von sechs allgemeinen Anforderungen an Datenqualitätsmetriken. Andererseits demonstriert ein reales Fallbeispiel die Instanziierbarkeit und Anwendbarkeit sowie den praktischen Mehrwert der neuen Metrik.

Schlüsselwörter: Datenqualität, Aktualität, Metrik, Messung

JEL-Classification: M15

Prof. Dr. Bernd Heinrich

Universität Regensburg – Lehrstuhl für Wirtschaftsinformatik II, Universitätsstraße 31,

93053 Regensburg, Deutschland

Email: Bernd.Heinrich@wiwi.uni-regensburg.de

Prof. Dr. Mathias Klier

Universität Regensburg – Professur für Wirtschaftsinformatik - Qualitätsmanagement und

Qualitätssicherung, Universitätsstraße 31, 93053 Regensburg, Deutschland

Email: Mathias.Klier@wiwi.uni-regensburg.de

Quirin Görz

Universität Augsburg – FIM Kernkompetenzzentrum Finanz- & Informationsmanagement,

Professur für Wirtschaftsinformatik und Management Support, Universitätsstraße 12, 86159

Augsburg, Deutschland

Email: Quirin.Goerz@wiwi.uni-augsburg.de

1. Einleitung

Qualitativ hochwertige Daten in Informationssystemen sind eine wichtige Grundlage für die Durchführung von Geschäfts-, Entscheidungs- und Unterstützungsprozessen (z. B. Al-Hakim 2007, S. 172, Ballou/Tayi 1999, S. 73). Dies gilt für unterschiedliche betriebswirtschaftliche Bereiche wie die Produktionsplanung und -kontrolle, das Supply Chain Management, das Kundenbeziehungsmanagement und das Controlling (Gustavsson/Wänström 2009, Kaplan et al. 1998, Kengpol 2006). So benötigen z. B. Produktionsplanungs- und -kontrollprozesse i. d. R. Daten aus einer Vielzahl unternehmensinterner und -externer Quellen (bspw. von Lieferanten oder Produktionspartnern), d. h. sie sind stark von der Qualität dieser Daten (z. B. deren Aktualität und Vollständigkeit) abhängig (Gustavsson/Wänström 2009). Monczka et al. (1998) zeigen zudem, dass erfolgreiche strategische Lieferantenbeziehungen in einem positiven Zusammenhang mit einer hohen Qualität der mit den Lieferanten ausgetauschten Daten stehen. Auch im Kundenbeziehungsmanagement hängt u. a. der Erfolg von Marketingkampagnen von der Qualität der verfügbaren Kundendaten ab. So gaben 76% der Befragten einer unter 500 Marketingleitern durchgeführten Studie an, dass die Qualität der Kundendaten einen direkten Einfluss auf die Profitabilität von Marketingkampagnen hat (SAS Institute 2006). Darüber hinaus spielen auch im Controlling die Datenqualität und deren Messung eine wichtige Rolle, um bspw. die Validität von Audits und Reports bestimmen zu können (Kaplan et al. 1998, S. 73).

Ist in diesen Fällen keine ausreichende Datenqualität gewährleistet, können Fehlentscheidungen und ein hoher Aufwand aus Datenqualitätsproblemen sowie deren Behebung resultieren (z. B. Ballou/Tayi 1999, S. 73, Even/Shankaranarayanan 2007, S. 75, Fisher et al. 2003, S. 170). So ergab eine Studie des Data Warehouse Institute, dass mangelhafte Datenqualität bei 67% der befragten Unternehmungen zu hohem Aufwand führt (Russom 2006, S. 11) (bspw. zur nachträglichen Fehlerbeseitigung in Unternehmensprozessen oder bei der Entwicklung neuer Informationssysteme). Ferner gaben 75% der Befragten einer internationalen Studie zum Thema Datenqualität an, dass bereits wichtige Entscheidungen aufgrund fehlerhafter Daten falsch getroffen wurden (Harris Interactive 2006). Die Sicherstellung der Vollständigkeit, Korrektheit und Aktualität von Daten, d. h. von Eigenschaften, die als Datenqualitätsdimensionen bekannt sind (Wang et al. 1995, S. 632), stellt somit für viele Unternehmungen ein relevantes Problem dar (vgl. z. B. Ballou et al. 1998, S. 462, Jiang et al. 2007, S. 1946, Russom 2006). Vor diesem Hintergrund beschäftigen sich auch zahlreiche wissenschaftliche Beiträge mit der Frage, wie Datenqualität in Informationssystemen zu messen ist (vgl. z. B. Ballou et al. 1998, Even/Shankaranarayanan 2007, Heinrich/Klier 2009, 2011, Heinrich et al. 2009, Lee et al. 2002, Parssian 2006, Parssian et al. 2004, 2009, Pipino et al. 2002).

Im Qualitätsmanagement sind mit Design- und Konformitätsqualität zwei unterschiedliche Perspektiven zu unterscheiden, die sich auch auf die Qualitätsmessung auswirken (Juran 1998). Designqualität bezeichnet den Grad der Übereinstimmung zwischen der Nachfrage der Datenanwender und der entsprechenden Repräsentation bspw. in einer Datenspezifikation (z. B. Datenschema). Hier existiert schon eine Vielzahl von Erhebungs- und Messverfahren, die v. a. der Informationsbedarfsanalyse zuzuordnen sind (vgl. z. B. Helfert 2002, Nicholas/Herman 2009, West 2011). Dagegen drückt die Konformitätsqualität aus, inwieweit die in einer Datenbank gespeicherten Datenattributwerte mit den zugehörigen Realweltausprägungen übereinstimmen (sind bspw. die gespeicherten Datenattributwerte noch korrekt?). Diese wird im Weiteren fokussiert, auch weil es hier im Gegensatz zur Informationsbedarfsanalyse wesentlich weniger wissenschaftliche Ansätze gibt. Die Unterscheidung der Qualitätsperspektiven ist im Hinblick auf die Messung der Datenqualität auch deshalb wichtig, weil dies die oftmals subjektive Datennachfrage und Einschätzung einzelner, befragter Anwender von der nachprüfbaren und reproduzierbaren Analyse der Konformität von tatsächlich gespeicherten Datenwerten und zugehörigen Realweltausprägungen abgrenzt.

Der Aktualität von Datenwerten kommt in der Wissenschaft eine besondere Bedeutung zu. So

stellt Aktualität eine der am stärksten diskutierten Datenqualitätsdimensionen dar (Al-Hakim 2007, S. 172, Klein/Callahan 2007, Lee et al. 2002, S. 134, Wand/Wang 1996). Dies liegt zum einen daran, dass Aktualität auf den zeitlichen Verfall gespeicherter *Datenwerte* fokussiert, was aus fachlicher/betriebswirtschaftlicher Perspektive bedeutend ist, gerade im Vergleich zu Dimensionen der (technischen) *Datenrepräsentation*, wie bspw. angemessenes Datenformat oder effiziente Speicherung. Zum anderen kommt der Aktualität aber auch im Vergleich zur verwandten Qualitätsdimension Korrektheit eine besondere Relevanz zu, da die Aktualität eine weitgehend automatisierbare Messung zu geringerem Aufwand verspricht und zudem eine Messung der Korrektheit vielfach sehr aufwendig oder nicht praktikabel ist (vgl. hierzu die Ausführungen in Abschnitt 2). Grundsätzlich werden bei der Messung der Qualitätsdimension Aktualität dabei primär Attributwerte fokussiert, die Zustände (in Unternehmensdatenbanken) dokumentieren, deren zeitlicher Verfall meist unternehmensexternen Einflüssen unterliegt und die vom Unternehmen selbst nicht wieder schnell und kostengünstig überprüft bzw. erneut bereitgestellt werden können.
Aber nicht nur wissenschaftliche Arbeiten betonen die Relevanz der Aktualität. Bspw. dokumentieren Analysen eines Datenbestands an Firmenkunden, dass Ansprechpartner in Firmen je nach Position mit einer Quote von 20-35% pro Jahr wechseln (Kraus 2004). Führt dies zu einer falschen oder erfolglosen Kundenansprache, kann erheblicher ökonomischer Schaden entstehen. Ähnliches ist bei Privatkunden festzustellen: Hier ergab die Untersuchung des Datenbestandes einer Unternehmung mit ca. 20 Millionen Kunden, dass pro Jahr ca. zwei Millionen Kunden umziehen, 230.000 sterben und 60.000 geschieden werden (Schönfeld 2007). Allein aufgrund der daraus resultierenden inaktuellen Daten entstand ein jährlicher Schaden bei der Kundenansprache von mehr als zwei Millionen Euro (Franz/von Mutius 2008).
Trotzdem existieren bisher in der Literatur keine Metriken zur Messung der Aktualität, die die wesentlichen Anforderungen im Rahmen eines ökonomisch orientierten Datenqualitätsmanagements erfüllen (vgl. Abschnitt 3). Dies ist alarmierend, zumal Metriken Ausgangspunkt dafür sind, ökonomische Auswirkungen einer schlechten bzw. verbesserten Datenqualität zu analysieren und effiziente Datenqualitätsmaßnahmen zu selektieren (vgl. Even et al. 2007, Heinrich/Klier 2006, Pipino et al. 2002, Shankaranarayanan/Cai 2006), und stellt aufgrund der Bedeutung der Aktualität eine Forschungslücke dar.
In der Arbeit wird daher diskutiert, wie basierend auf einem konformitätsorientierten Qualitätsverständnis die Aktualität von Daten in Informationssystemen mittels einer Metrik gemessen werden kann. Zur Bearbeitung dieser Fragestellung wird eine normative, quantitative Modellierung (vgl. Bertrand/Fransoo 2002, Meredith et al. 1989) verfolgt, d. h. es wird ein gestaltungsorientierter Beitrag zur Entwicklung einer Datenqualitätsmetrik für Aktualität angestrebt. Die Arbeit ist wie folgt strukturiert: In Abschnitt 2 werden die betrachtete Problemstellung konkretisiert und der Untersuchungsgegenstand eingegrenzt. In Abschnitt 3 werden sechs Anforderungen an Datenqualitätsmetriken aufgegriffen, darauf basierend Stärken und Schwächen bestehender Aktualitätsmetriken diskutiert und resultierender Forschungsbedarf identifiziert. Abschnitt 4 beinhaltet die Konstruktion einer neuen Metrik für Aktualität, die auf Grundlagen der Wahrscheinlichkeitstheorie basiert. In Abschnitt 5 wird zunächst die neue Metrik im Hinblick auf die sechs Anforderungen und den identifizierten Forschungsbedarf evaluiert. Danach wird die Metrik für das Fallbeispiel eines großen Mobilfunkanbieters instanziiert, um ihre Anwendbarkeit und den praktischen Mehrwert zu demonstrieren. Der letzte Abschnitt fasst den Beitrag zusammen und würdigt diesen kritisch.

2. Problembeschreibung und Forschungsgegenstand

In der Literatur existiert eine Reihe von Definitionen für den Begriff Datenqualität. Laut Hinrichs (2002, S. 26) stellt Datenqualität den „Grad, in dem ein Satz inhärenter Merkmale eines Datenprodukts Anforderungen erfüllt“, dar. Orr (1998, S. 67) definiert, dass „data quality is the measure of the agreement between the data views presented by an information system and

that same data in the real world“ und Parrsian et al. (2004, S. 967) halten fest, dass „the terms information quality and data quality have been used to characterize mismatches between the view of the world provided by an IS and the true state of the world“. Zusammenfassend wird unter Datenqualität hier im Wesentlichen die Eigenschaft verstanden, inwieweit die in einem Informationssystem gespeicherten Daten definierten Erfordernissen genügen bzw. den zugehörigen Realweltausprägungen entsprechen. Wie bereits oben motiviert, wird dabei im Folgenden die Konformitätsqualität fokussiert. Demnach ist Datenqualität in Bezug auf die tatsächlich gespeicherten Datenwerte und die zugehörigen Realweltausprägungen zu ermitteln (Analyse der Konformität). Als Problemstellung wird in der vorliegenden Arbeit die konformitätsorientierte Messung der Aktualität betrachtet.
Zur weiteren Konkretisierung des Forschungsgegenstands wird zunächst die Dimension Korrektheit kurz erörtert (vgl. Tab. 1). Danach wird die Dimension Aktualität definiert und davon abgegrenzt.

Autoren	Begriff und Definition
Ballou/Pazer (1985, S. 153)	Accuracy: „The recorded value is in conformity with the actual value.“
Wang/Strong (1996, S. 31)	Accuracy: „The extent to which data are correct, reliable, and certified free of error.“
Redman (1996, S. 255)	Accuracy: „Accuracy of a datum <*e*, *a*, *v*> refers to the nearness of the value *v* to some value *v*' in the attribute domain, which is considered as the correct one for the entity *e* and the attribute *a*.“
Hinrichs (2002, S. 30)	Korrektheit: „die Eigenschaft, dass die Attributwerte eines Datenprodukts (im Informationssystem) denen der modellierten Entitäten (in der Diskurswelt) entsprechen.“
Eppler (2003, S. 77)	Accuracy: „how closely information matches a real-life state.“
Batini/Scannapieco (2006, S. 20)	Accuracy: „Accuracy is defined as the closeness between a value *v* and *a* value *v*', considered as the correct representation of the real-life phenomenon that *v* aims to represent.“

Tab. 1 Definitionen für Korrektheit (Auswahl)

Von vielen Autoren wird Korrektheit als Nähe eines Attributwerts ω zur zugehörigen Realweltausprägung verstanden (in welchem Ausmaß stimmt ein Attributwert ω mit den realen Gegebenheiten überein?). Die Messung erfolgt hier durch einen Vergleich des Attributwerts ω mit der zugehörigen Realweltausprägung, was u. a. auch Wang/Strong (1996) explizit anführen.
Basierend auf der Begriffsdefinition für Korrektheit wird die Definition für Aktualität spezifiziert. Hier werden anhand der Tab. 2 auch wichtige Unterschiede zwischen diesen Dimensionen deutlich (mit timeliness, currency und Zeitnähe sind auch in der Literatur synonym verwendete Begriffe berücksichtigt).

Autoren	Begriff und Definition
Ballou/Pazer (1985, S. 153)	Timeliness: „The record value is not out of date.“
Wang/Strong (1996, S. 32)	Timeliness: „The extent to which the age of the data is appropriate for the task at hand.“
Redman (1996, S. 258)	Currency: „Currency refers to a degree to which a datum in question is up-to-date. A datum value is up-to-date if it is correct in spite of possible discrepancies caused by time-related changes to the correct value.“
Hinrichs (2002, S. 31)	Zeitnähe: „die Eigenschaft, dass die Attribute beziehungsweise Tupel eines Datenprodukts jeweils dem aktuellen Diskursweltzustand entsprechen, das heißt nicht veraltet sind.“
Batini/Scannapieco (2006, S. 29)	Timeliness: „Timeliness expresses how current data are for the task at hand.“

Tab. 2 Definitionen für Aktualität (Auswahl)

Im Kern wird Aktualität als die Eigenschaft eines Attributwerts ω verstanden, dass ein in der Vergangenheit korrekt erfasster Attributwert ω zum Bewertungszeitpunkt (Zeitpunkt der Messung) nach wie vor den realen Gegebenheiten entspricht und inzwischen nicht veraltet und somit inkorrekt ist (vgl. insbesondere Redman 1996, S. 258). Im Gegensatz zur Korrektheit stellt Aktualität somit auf den zeitlichen Verfall eines (gespeicherten) Attributwerts ω ab. Dies setzt voraus, dass der Attributwert ω in der Vergangenheit entsprechend der realen Ge-

gebenheiten (Realweltabgleich) ein oder mehrmalig erfasst bzw. überprüft wurde. Beim letzten Erfassungszeitpunkt eines Attributwerts ω kann es sich demnach sowohl um die erstmalige Erfassung als auch um einen darauffolgenden Abgleich des Attributwerts ω (dieser wird bestätigt oder geändert) handeln. Zum Bewertungszeitpunkt kann schließlich (alternativ) eine Messung der Korrektheit (d. h. eine erneuter Realweltabgleich) oder der Aktualität des Attributwerts ω erfolgen. Anhand dieser Unterschiede lässt sich das jeweilige Verständnis der beiden hier betrachteten Dimensionen noch einmal verdeutlichen (vgl. Abb. 1).

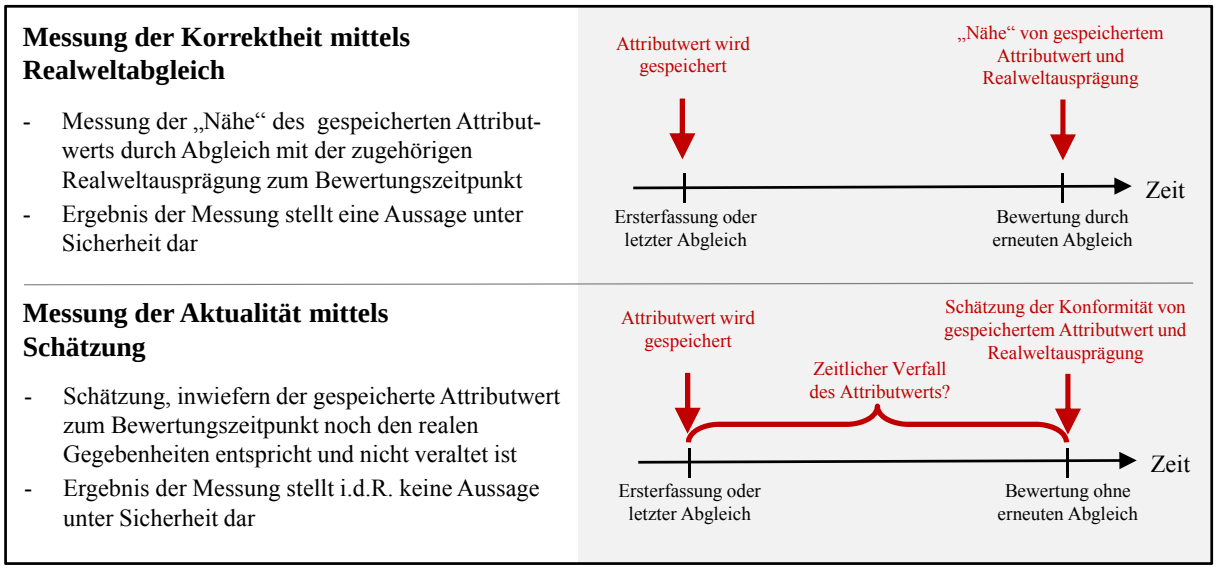

Abb. 1 Messung der Korrektheit vs. der Aktualität

Bei der Messung der Korrektheit zum Bewertungszeitpunkt erfolgt ein Realweltabgleich. Dabei wird zum Bewertungszeitpunkt die Nähe des betrachteten Attributwerts ω zur zugehörigen Realweltausprägung im Allgemeinen mittels eines Ähnlichkeits- oder Distanzmaßes (z. B. Hamming-Distanz) quantifiziert (vgl. Heinrich et al. 2008, Hinrichs 2002). Das Ergebnis der Messung der Korrektheit stellt daher eine Aussage unter Sicherheit dar. Im Gegensatz zur Korrektheit wird bei der Messung der Aktualität kein expliziter Realweltabgleich gefordert, d. h. kein direkter Vergleich des Attributwerts ω mit der zugehörigen Realweltausprägung. Stattdessen wird mithilfe einer *Schätzung* ermittelt, inwiefern ein Attributwert ω zum Bewertungszeitpunkt noch immer den realen Gegebenheiten entspricht und seit dem Zeitpunkt der korrekten Erfassung nicht (bedingt durch zeitlichen Verfall) veraltet ist. Das Ergebnis der Messung der Aktualität stellt daher i. d. R. keine Aussage unter Sicherheit dar. Folglich lässt sich die Bedeutung der Dimension Aktualität damit begründen, dass ein wiederholt erforderlicher Realweltabgleich (wie bei der Messung der Korrektheit) oftmals sehr aufwendig oder nicht praktikabel ist (vgl. das spätere Fallbeispiel). Dennoch soll eine Aussage über die Qualität des Attributwerts ω möglich sein.

Soll demnach ein aufwendiger Realweltabgleich vermieden werden, stellt sich die Frage, welche Daten für die Messung der Aktualität eines Attributwerts ω (z. B. Attributwert „Student“ in einer Kundendatenbank) heranzuziehen sind. Die Autoren in Tab. 2 führen hier primär *attributwertspezifische Metadaten* an. Beispiele dafür sind der Entstehungszeitpunkt t_0 der zugehörigen Realweltausprägung und die Gültigkeitsdauer T dieser Ausprägung. Je nachdem, ob diese Metadaten zum Bewertungszeitpunkt t_1 bekannt sind oder nicht, erfolgt die Messung der Aktualität eines Attributwerts ω unter Sicherheit oder Unsicherheit. Gemäß der obigen Definition ist zu prüfen, ob der betrachtete Attributwert ω zum Bewertungszeitpunkt t_1 nach wie vor den realen Gegebenheiten entspricht. D. h., es ist zu ermitteln bzw. abzuschätzen, ob $(t_1\text{-}t_0)\leq T$ gilt.

Im hypothetischen Fall, dass der Entstehungszeitpunkt t_0 und die Gültigkeitsdauer T der zugehörigen Realweltausprägung bekannt sind, könnte unter Sicherheit bestimmt werden, inwiefern der betrachtete Attributwert ω zum Bewertungszeitpunkt t_1 noch den realen

Gegebenheiten entspricht. Wenn $(t_1-t_0)>T$ gilt, ist der Attributwert demnach inkorrekt. Andernfalls entspricht der Attributwert zum Bewertungszeitpunkt definitiv noch der zugehörigen Realweltausprägung ($(t_1-t_0)\leq T$) und ist korrekt. Im Gegensatz zu dieser (trivialen) Messung der Aktualität unter Sicherheit wird im Weiteren diejenige unter Unsicherheit fokussiert. Diese ist wesentlich realistischer, da die Gültigkeitsdauer T eines Attributwerts ω zum Bewertungszeitpunkt zumeist nicht bekannt ist. Somit stellt sich die Frage, wie eine Messung der Aktualität dennoch erfolgen kann, wenn Metadaten nicht oder nur zum Teil bekannt sind.
Hier sind weitere Attributwerte w_i (i=1,…,n) anzuführen, die in einem Zusammenhang bspw. mit der unbekannten Gültigkeitsdauer T des Attributwerts ω stehen und Rückschlüsse auf diese zulassen (vgl. hierzu auch Heinrich/Klier 2009). Sie werden als *Zusatzdaten* bezeichnet. Ihre Bedeutung wird kurz am Beispiel des Datenwerts „Student“ erläutert. Abb. 2 illustriert den Zusammenhang zwischen der Studiendauer (unter Berücksichtigung der Studienabbrecher), d. h. der Gültigkeitsdauer T des Werts „Student“, und der *Hochschulart* (hier *Universität* versus *Fachhochschule*)[1]. So entspricht bspw. der Wert „Student“ bei Personen, die sich an einer Fachhochschule für ein ingenieurwissenschaftliches Studienfach eingeschrieben haben (ohne weitere Daten zu berücksichtigen[2]), nach elf Semestern nur noch bei 18% den realen Gegebenheiten. Dieser Wert errechnet sich dabei als Differenz aus eins und der kumulierten relativen Häufigkeit der Studienabgänger (erfolgreicher Abschluss sowie Abbruch – vgl. rechte Seite in Abb. 2) nach elf Semestern (hier 1-0,82). Im Gegensatz dazu entspricht der Wert „Student“ bei der *Hochschulart* Universität und der *Studienfachgruppe* Ingenieurwissenschaften immerhin noch in 46% (d. h. erst 54% der initial eingeschriebenen Studenten sind bereits Studienabgänger – vgl. linke Seite in Abb. 2) der Fälle den realen Gegebenheiten. Folglich sind Zusatzdaten wie bspw. die *Hochschulart* relevant für die Schätzung der Aktualität des Datenwerts „Student“.

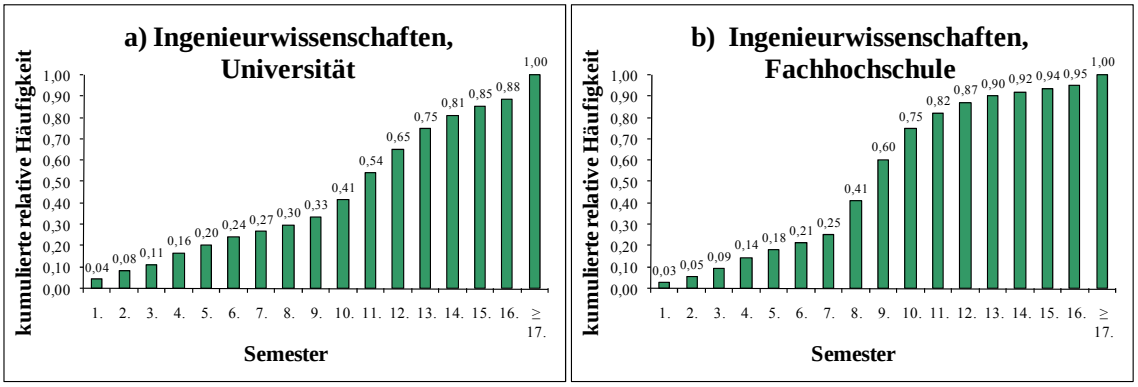

Abb. 2 Kumulierte relative Häufigkeiten von Studiendauern (inkl. Abbrecher)

Zusammenfassend sind bei der Messung der Aktualität neben attributwertspezifischen Metadaten auch Zusatzdaten zu berücksichtigen. Sind Zusatzdaten neben dem Attributwert ω bereits als weitere Attributwerte w_i in der Datenbank gespeichert (vgl. auch das Fallbeispiel in Abschnitt 5.2), so versprechen sie eine verbesserte Güte der Qualitätsmessung ohne zusätzlichen Erhebungsaufwand.

3. Anforderungen an Datenqualitätsmetriken und existierende Aktualitätsmetriken

Im Folgenden werden existierende Aktualitätsmetriken analysiert. Um diese Analyse in Abschnitt 3.2 strukturiert durchzuführen, werden in Abschnitt 3.1 zunächst sechs Anforderungen R.1 bis R.6 an datenwertorientierte Qualitätsmetriken aus der Literatur entnommen, konkretisiert und aus Sicht eines ökonomisch orientierten Datenqualitätsmanagements begründet.

3.1. Anforderungen an Datenqualitätsmetriken

Wie in der Einleitung beschrieben, dienen Datenqualitätsmetriken als eine Basis zur Realisierung eines ökonomisch orientierten Datenqualitätsmanagements. Dieser Zusammenhang wird in der Literatur oftmals mithilfe eines Datenqualitätsregelkreises dargestellt (Feigenbaum 1991, S. 316 ff., Heinrich et al. 2009, S. 23 ff.). Abb. 3 zeigt einen solchen, vereinfachten Datenqualitätsregelkreis und veranschaulicht, wie die Messung des vorhandenen Datenqualitätsniveaus anhand einer Metrik einen zielorientierten Einsatz von Qualitätsmaßnahmen (wie z. B. Datenbereinigung oder Zukauf externer Daten) erlauben soll. Hierbei stellt sich die Frage, in welchem ökonomisch sinnvollen Umfang derartige Maßnahmen (Regler) zur Qualitätsverbesserung zu ergreifen sind. Dieser Umfang ist – basierend auf der Messung des vorhandenen Datenqualitätsniveaus – unter ökonomischen Gesichtspunkten festzulegen. Über die Messung des ex post realisierten Datenqualitätsniveaus kann anschließend die Effektivität einer durchgeführten Maßnahme beurteilt werden.

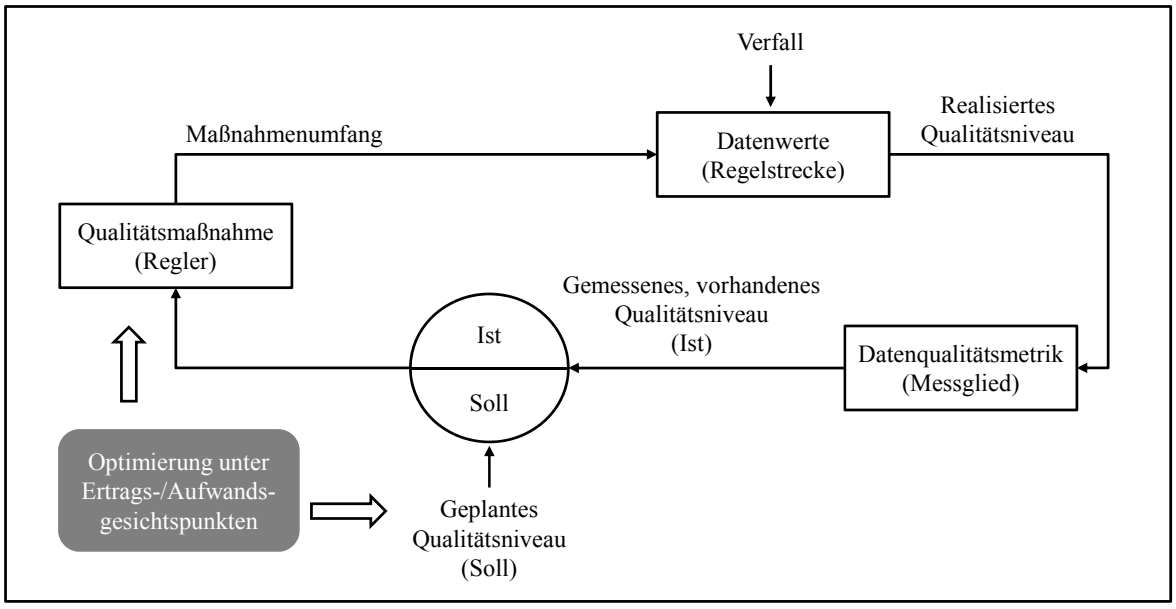

Abb. 3 Vereinfachter Datenqualitätsregelkreis

Damit eine Datenqualitätsmetrik diese Aufgabe im Regelkreis übernehmen kann und auch, um eine Fundierung und Nachvollziehbarkeit bei der Konstruktion der neuen Metrik in Abschnitt 4 zu gewährleisten, werden im Folgenden die Anforderungen R.1 bis R.6 an datenwertorientierte Qualitätsmetriken zugrunde gelegt. Die Anforderungen wurden aus der Literatur entnommen (vgl. Ballou et al. 1998, S. 467 ff., Even/Shankaranarayanan 2007, S. 83 ff., Heinrich et al. 2009, S. 5:4 f., Pipino et al. 2002, S. 213), konkretisiert und aus Sicht eines ökonomisch orientierten Datenqualitätsmanagements begründet:

R.1 [*Normierung*] Damit eindeutig definiert ist, wann ein Datenwert die maximale bzw. minimale Qualität hinsichtlich einer Qualitätsdimension aufweist (z. B. wann gilt ein Datenwert definitiv als aktuell bzw. inaktuell), sind die Metrikergebnisse geeignet zu normieren. D. h., das Metrikergebnis für einen beliebigen Datenwert muss das Resultat einer Abbildung auf einen Wertebereich mit unterer und oberer Schranke sein (etwa Infimum von null und Supremum von eins bei einem Wertebereich von [0; 1]).

Wird eine solche Normierung auf einen beschränkten Wertebereich nicht gewährleistet, ist nicht eindeutig entscheidbar, bei welchem Metrikergebnis ein Datenwert als definitiv aktuell (Supremum) bzw. inaktuell (Infimum) gilt. Im Rahmen eines ökonomisch orientierten Datenqualitätsmanagements reduziert dies, gerade bei wiederholten Messungen im Zeitverlauf, die Aussagekraft der Metrikergebnisse und kann zu falschen Entscheidungen führen.

R.2 [*Kardinalskalierung*] Um eine exakte Vergleichbarkeit verschiedener Metrikergebnisse zu gewährleisten, ist deren Kardinalskalierung zu fordern. Eine Kardinalskala liegt vor, „wenn die Ausprägungen des untersuchten Merkmals nicht nur in eine Rangordnung

gebracht werden können, sondern zusätzlich noch bestimmt werden kann, in welchem Ausmaß sich je zwei verschiedene Merkmalsausprägungen unterscheiden“ (Bamberg et al. 2007, S. 7). Die Metrik muss daher die Eigenschaft aufweisen, dass Differenzen zwischen Metrikergebnissen exakt bestimmbar und aussagekräftig sind (Cramer/Kamps 2008, S. 8, Fahrmeir et al. 2010, S. 18).

Neben der Normierung ist auch die Kardinalskalierung der Metrikergebnisse für ein ökonomisch orientiertes Datenqualitätsmanagement notwendig. Sind bspw. zwei Maßnahmen nicht nur hinsichtlich ihrer Aufwands-, sondern auch ihrer Qualitätswirkung zu vergleichen, so reicht die Aussage nicht aus, dass beide Maßnahmen die Datenqualität verbessern bzw. welche stärker zur Verbesserung beiträgt (für den Fall einer Ordinalskalierung). Vielmehr ist es zur Selektion der ökonomisch sinnvollen Maßnahmen in einem ersten Schritt nötig, das *Ausmaß* der jeweiligen Verbesserung und damit die *Differenzen* zwischen Metrikergebnissen exakt zu bestimmen und dem Maßnahmenaufwand gegenüber zu stellen.

R.3 [*Interpretierbarkeit*] Damit die Metrikergebnisse eindeutig interpretierbar sind, müssen diese eine Dimension oder Maßeinheit besitzen (Lindner et al. 2006, S. 24 ff.). So ist mindestens die Maßeinheit eins zu fordern (wie z. B. bei Anzahlen, Wahrscheinlichkeiten und Verhältniszahlen, d. h. Quotienten aus Größen mit gleicher Maßeinheit) (Bureau International des Poids et Mesures 2006, S. 120). Bspw. besitzen die Metrikergebnisse für die Qualitätsdimension Vollständigkeit dann die Maßeinheit eins, wenn diese als Verhältnis der Anzahl tatsächlich gespeicherter Datenwerte zur Anzahl der zugehörigen Realweltausprägungen gemessen werden, da beide Größen die gleiche Maßeinheit besitzen.

Anforderung R.3 unterstützt den zweiten Schritt zur Ermittlung der ökonomisch sinnvollen Qualitätsmaßnahmen. Ziel ist hier, die Qualitätsverbesserung in eine monetäre Einheit zu überführen und dadurch einen direkten Vergleich mit dem Maßnahmenaufwand zu gewährleisten. Dafür sind Metrikergebnisse (Repräsentation der Qualitätsverbesserung) mit einer Dimension oder Maßeinheit notwendig. Nur dann sind die Bestimmung und v. a. die Interpretation der (Differenzen unterschiedlicher) Metrikergebnisse eindeutig.

R.4 [*Aggregierbarkeit*] Metrikergebnisse für einzelne Attributwerte müssen zu konsistenten Ergebnissen für eine Menge an Attributwerten aggregierbar sein. So muss bspw. bei Zugrundelegung eines relationalen Datenbankmodells das Datenqualitätsniveau auf Attributwert-, Tupel-, Relationen- sowie Datenbankebene zueinander konsistent ermittelbar sein. Hierzu sind Aggregationsvorschriften zu fordern, die für die Metrikergebnisse der unterschiedlichen Ebenen jeweils eine konsistente Interpretation (z. B. bei Vollständigkeit jeweils als Anteil der gespeicherten Datenwerte im Verhältnis zu den zugehörigen Realweltausprägungen) sowie Repräsentation (z. B. jeweils als Resultat einer Abbildung auf den Wertebereich [0; 1]) gewährleisten.

Anforderung R.4 zielt darauf ab, dass sich das Datenqualitätsmanagement realistischerweise nicht nur auf die isolierte Messung der Datenqualität eines Datenattributs beschränkt. Vielmehr ist die Datenqualität von Mengen an Datenattributen zu betrachten, die bspw. in Tupeln, Relationen oder Views organisiert sind. Insofern ist es notwendig, dass die Metrikergebnisse für eine beliebige Menge an Datenattributen konsistent aggregierbar sind. Existieren keine Aggregationsvorschriften bzw. gewährleisten definierte Aggregationsvorschriften eine konsistente Interpretation sowie Repräsentation der Metrikergebnisse nicht, bleibt das Datenqualitätsmanagement auf die Partialsicht einzelner Datenwerte beschränkt.

R.5 [*Konfigurierbarkeit*] Zur anwendungskontextspezifischen (nicht nutzerspezifischen) Messung der Datenqualität ist die Konfigurierbarkeit der Metrik für verschiedene Domänen zu fordern. Dies bedeutet:

a) [*Gewichtung*] Die Metrik muss eine Gewichtung der Attribute hinsichtlich deren Bedeutung im Anwendungskontext erlauben. Demnach muss es zur Qualitätsmes-

sung auch möglich sein, die Metrik nur auf eine Teilmenge der Datenattribute anzuwenden (nicht relevante Datenattribute erhalten die Gewichtung null).

b) [*Attributspezifische Charakteristika*] Die Metrik muss so konfigurierbar sein, dass sie den spezifischen Charakteristika der Datenattribute (z. B. in Bezug auf die Dimension Aktualität, den unterschiedlichen zeitlichen Verfallraten) entsprechend Rechnung trägt.

c) [*Berücksichtigung von Zusatzdaten*] Die Metrik muss bei der Messung der Datenqualität eine Berücksichtigung von Zusatzdaten[3] gewährleisten.

Anforderung R.5 stellt auf die Gewährleistung eines möglichst breiten Einsatzgebietes einer Metrik ab. Gilt es bspw., die Datenqualität für verschiedene Datentupel zu messen, wobei im Anwendungskontext bestimmte Attribute der Tupel irrelevant sind (bspw. die Telefonnummer des Kunden im Hinblick auf eine postalisch durchzuführende Marketingkampagne), so dürfen diese Attribute bei der Berechnung der Metrikergebnisse nicht eingehen. Auch können Probleme resultieren, falls den spezifischen Charakteristika von Attributen (bspw. den Verfallraten bei der Dimension Aktualität) in der Metrik nicht Rechnung getragen wird. Ebenso ist die Berücksichtigung von Zusatzdaten bei der Qualitätsmessung relevant.

R.6 [*Operationalisierbarkeit*] Zur Gewährleistung der Operationalisierbarkeit und Anwendbarkeit der Metrik ist Folgendes zu fordern:

a) [*Ermittelbarkeit der Inputgrößen*] Die Inputgrößen der Metrik müssen ermittelbar sein.

b) [*Automatisierbarkeit*] Die Metrikergebnisse sind automatisiert zu berechnen.

Anforderung R.6 zielt ebenfalls auf ein breites Einsatzgebiet einer Metrik ab. Sind die Inputgrößen einer Metrik nicht oder nur mit sehr hohem Aufwand ermittelbar, so hat die Metrik allenfalls theoretische Bedeutung. Ähnliches gilt, falls die Berechnung der Metrikergebnisse nicht automatisierbar ist. So ist realistischerweise davon auszugehen, dass eine Metrik i. d. R. nicht nur dazu eingesetzt wird, die Qualität einiger weniger Datenwerte zu messen. Soll demnach die Qualität größerer Datenmengen geprüft werden, ist eine manuelle Berechnung der Ergebnisse zu vermeiden.

Insgesamt soll mit den obigen Anforderungen R.1 bis R.6 eine begründete, notwendige Basis zur Realisierung eines ökonomisch orientierten Datenqualitätsmanagements und damit auch zur Analyse existierender Aktualitätsmetriken gelegt werden.

3.2. Analyse existierender Aktualitätsmetriken

Im Folgenden werden mithilfe der Anforderungen R.1 bis R.6 bestehende Metriken zur Messung der Aktualität analysiert. Auf diese Weise ist es möglich, den Stand der Forschung zielgerichtet und nachvollziehbar aufzuarbeiten. Hierbei wird jedoch ausdrücklich nicht darauf abgezielt, bestehende Metriken zu kritisieren. Vielmehr sollen konstruktiv Forschungslücken identifiziert werden.

Um eine konsistente Analyse durchführen zu können, muss den untersuchten Ansätzen eine konformitätsorientierte Datenqualitätsperspektive zugrunde liegen. Zudem müssen die Metriken eindeutig (z. B. formal) definiert oder anhand der Ausführungen eindeutig definierbar sein, um die Nachvollziehbarkeit der Analyse zu gewährleisten. Hierunter fallen die Ansätze von Hinrichs (2002), Ballou et al. (1998) sowie Even/Shankaranarayanan (2007). Für andere, wie z. B. die AIM Quality-Methode (vgl. Lee et al. 2002), die Total Quality Data Methodology (vgl. English 1999) oder den prozessorientierten Ansatz von Redman (1996) gilt dies nicht, da hier keine Metriken formal definiert werden. Ähnliches ist auch für den Ansatz von Heinrich et al. (2009) festzuhalten, die nicht die Entwicklung einer formal definierten Metrik für Aktualität anstreben, sondern ein allgemeines Vorgehen zur Entwicklung von Metriken vorschlagen.

Der Ansatz von Hinrichs (2002) liefert zum Bewertungszeitpunkt der Aktualität eine Schätzung dafür, dass ein betrachteter Attributwert noch nicht veraltet ist. Die Metrik für Aktualität wird mittels des folgenden, in der Notation leicht modifizierten, Quotienten angegeben:

$$Q^{\omega}_{Akt.}(t', A) = \frac{1}{Upd(A) \cdot t' + 1} \tag{1}$$

Dabei repräsentiert $t' \in R_0^+$ die Speicherdauer des Attributwerts ω, die sich als Differenz aus dem Bewertungszeitpunkt t_1 und dem Erfassungszeitpunkt t_0' des Attributwerts ω, d. h. dem Zeitpunkt der Erfassung der zugehörigen Realweltausprägung, ergibt. Der Erfassungszeitpunkt t_0' wird dabei ebenso als bekannt angenommen wie die mittlere Änderungshäufigkeit[4] von Werten des Attributs A. Letztere wird in Term (1) durch $Upd(A) \in R_0^+$ repräsentiert und gibt an, wie oft Werte des Attributs A sich im Durchschnitt innerhalb einer definierten Zeitperiode in der Realwelt ändern (z. B. zehnmal pro Jahr).
Mit Bezug zu den Anforderungen: Ist die Änderungshäufigkeit $Upd(A)$ oder die Speicherdauer t' des Attributwerts ω gleich null, so ist der Attributwert aktuell und das Metrikergebnis $Q^{\omega}_{Akt.}(t', A)$ ergibt sich zu eins (maximaler Wert). Steigt dagegen die Speicherdauer t' des Attributwerts ω oder die Änderungshäufigkeit $Upd(A)$ an, verringert sich ceteris paribus das Metrikergebnis. Gilt dabei ($Upd(A) \to \infty$ und $t' > 0$) oder ($Upd(A) > 0$ und $t' \to \infty$), so folgt $Q^{\omega}_{Akt.}(t', A) \to 0$ (minimaler Wert). Das Metrikergebnis für einen beliebigen Attributwert ist insofern das Resultat einer Abbildung auf einen beschränkten, zusammenhängenden Wertebereich (R.1). Dagegen sind die Metrikergebnisse aufgrund des Quotienten weder kardinalskaliert noch interpretierbar im Sinne der Anforderungen R.2 und R.3. Werden zwei Metrikergebnisse (bspw. gemessen zu zwei unterschiedlichen Zeitpunkten) für einen Attributwert verglichen und ist der zu einem früheren Zeitpunkt gemessene Wert größer als der später gemessene Wert, so entspricht dies der Einschätzung, dass sich die Aktualität des betrachteten Attributwerts verringert hat. Diese Einschätzung ist jedoch auf das Festlegen einer ordinalen Rangfolge beschränkt. Die Differenz zweier Metrikergebnisse (was bedeutet eine Reduzierung des Metrikergebnisses um z. B. 0,1?) hat keine weitere Aussagekraft. Auch besitzen die resultierenden Werte dieser Metrik sowie entsprechende Differenzen keine Dimension oder Maßeinheit und sind damit nicht interpretierbar. Für die Aggregation der Metrikergebnisse auf Tupel-, Relationen- und Datenbankebene sind Vorschriften definiert, jedoch fehlen auf Datenbankebene Parameter, um eine konsistente Interpretation auf allen Ebenen zu gewährleisten (R.4). Das hier zugrunde gelegte ungewichtete arithmetische Mittel hat zur Folge, dass Relationen mit einer größeren Anzahl enthaltener Attribute und Tupel sowie einer größeren Bedeutung im Anwendungskontext bei Bedarf nicht stärker gewichtet werden können. Zudem hängt der Wert der Metrik auf Datenbankebene somit von der Zerlegung der Datenbank in Relationen ab. Insofern führen die Aggregationsvorschriften nur für den Einzelfall, dass die Zerlegung in Relationen deren Bedeutung im Anwendungskontext exakt widerspiegelt, zu einer konsistenten Interpretation. Dagegen ist eine kontextspezifische Konfiguration der Metrik im Sinne einer Gewichtung (mit Ausnahme der Datenbankebene) möglich (R.5a)). Die in R.5b) beziehungsweise R.5c) geforderte Berücksichtigung von attributspezifischen Charakteristika respektive Zusatzdaten ist nicht gewährleistet. So sind die Werte vieler realer Attribute (wie z. B. Familienstand und Berufsstatus) durch eine sich im Zeitverlauf verändernde Verfallrate charakterisiert (vgl. hierzu insbesondere auch die Untersuchungen zum Attributwert „Student“ im Rahmen des in Abschnitt 5.2 dargestellten Fallbeispiels). Dies lässt sich mit der obigen Metrik ebenso nicht berücksichtigen, wie Zusatzdaten. Die Speicherdauer t' kann bei bekanntem Erfassungszeitpunkt t_0' bestimmt werden, wobei herkömmliche Datenbanken diesen Zeitpunkt gewöhnlich als Metadatum speichern. Die Änderungshäufigkeit $Upd(A)$ eines Attributs A kann geschätzt oder aufgrund von Erfahrungswerten (historische Daten) festgelegt werden. Die Ermittelbarkeit der Inputgrößen laut R.6a) ist

damit gewährleistet. Liegen diese Größen für ein Attribut bzw. die Attributwerte vor, ist die Berechnung der Metrikergebnisse automatisierbar (R.6b)).
Ballou et al. (1998) schlagen folgende, in der Notation leicht modifizierte, Metrik vor:

$$Q^{\omega}_{Akt.}(t, A) = \left[\max\left\{ 1 - \frac{t}{T_{\max}(A)}; 0 \right\} \right]^{s} \tag{2}$$

Das Alter $t \in R_0^+$ eines Attributwerts ω ist als Differenz zwischen dem Bewertungszeitpunkt t_1 und dem als bekannt angenommenen Entstehungszeitpunkt t_0 der zugehörigen Realweltausprägung definiert. Der Parameter $T_{\max}(A) \in R^+$ entspricht der ebenfalls als bekannt angenommenen, *maximalen* Gültigkeitsdauer von Werten des Attributs A. Eine Zunahme der maximalen Gültigkeitsdauer $T_{\max}(A)$ führt ceteris paribus zu einem höheren Metrikergebnis et vice versa. Hier ist jedoch anzumerken, dass für viele Attribute eine solche *feste* maximale Gültigkeitsdauer $T_{\max}(A)$ nicht bekannt ist oder nicht existiert. Die Auswirkung des Quotienten aus t und $T_{\max}(A)$ auf das Metrikergebnis kann kontextspezifisch durch den Exponenten s beeinflusst werden. Die Metrik von Ballou et al. (1998) ist infolge der Maximumfunktion für $s \in R^+$ auf das Intervall [0; 1] normiert. D. h., das Metrikergebnis für einen beliebigen Attributwert ist das Resultat einer Abbildung auf das beschränkte, zusammenhängende Intervall [0; 1] (R.1). Eine Kardinalskalierung der Metrikergebnisse ist nur im Einzelfall s=1 gewährleistet (R.2). In diesem Fall würde jedoch die kontextspezifische Konfigurierbarkeit mithilfe des Parameters s entfallen, die Ballou et al. (1998) explizit als Vorteil der Metrik anführen. In Bezug auf die Interpretierbarkeit der Metrikergebnisse wird weder eine Dimension oder Maßeinheit vorgeschlagen, noch ist eine solche abzuleiten. Nur wiederum für s=1 und in Verbindung mit einer im Intervall [0; $T_{\max}(A)$] gleichverteilten Gültigkeitsdauer T der betrachteten Attributwerte ω ist das Metrikergebnis als Wahrscheinlichkeit zu interpretieren. Dieser Einzelfall wird jedoch von den Autoren nicht diskutiert. Wird nämlich s=1 gesetzt, schränkt dies die Einsatzgebiete der Metrik sehr ein, da damit die Annahmen einer festen maximalen Gültigkeitsdauer der Attributwerte $T_{\max}(A)$ und einer konstanten absoluten Verfallrate einhergehen. Diese Annahmen treffen für viele Attribute, wie z. B. *Adresse*, *Nachname* und *Berufsstatus*, nicht zu. Für $s \neq 1$ lässt sich dagegen zeigen, dass die Metrikergebnisse keine Maßeinheit und damit keine Interpretation besitzen (R.3). Für s=1 lässt sich nämlich das Metrikergebnis als Anteil (z. B. Prozentsatz) der restlichen Gültigkeitsdauer (maximale Gültigkeitsdauer $T_{\max}$ abzüglich des Alters t eines Attributwerts ω beides gemessen in der gleichen zeitlichen Maßeinheit) an der maximalen Gültigkeitsdauer noch ohne Weiteres interpretieren. Durch die Potenzierung mit dem Exponenten s für $s \neq 1$ geht allerdings diese Interpretierbarkeit verloren. So wird bspw. für s=2 die resultierende Differenz (bspw. 0,4 falls das Alter t eines Attributwerts ω bereits 60% seiner maximalen Gültigkeitsdauer erreicht hat) quadriert, was zur Folge hat, dass die bisherige Interpretation als Anteil (es würde $0{,}4^2$=0,16 resultieren) nicht mehr zulässig ist. Eine eindeutige Interpretation der Metrikergebnisse im Sinne einer Maßeinheit ist in diesem Fall nicht mehr gewährleistet. Aggregationsvorschriften im Sinne von R.4 werden nicht vorgeschlagen. Dagegen kann die Metrik durch Wahl des Exponenten s anwendungskontextspezifisch konfiguriert werden (R.5a)). Attributspezifische Charakteristika und Zusatzdaten (R.5b) und R.5c)) werden nicht betrachtet. Hier ist es bspw. nicht möglich, sich im Zeitverlauf verändernde Verfallraten abzubilden. Bezüglich R.6a) ist festzustellen, dass viele Attribute über keine *feste maximale* Gültigkeitsdauer der Attributwerte verfügen (z. B. Gültigkeitsdauer einer Kundenadresse). Dies schränkt die Ermittelbarkeit der Inputgrößen der Metrik ein. Zudem muss zur automatisierbaren Berechnung der Metrikergebnisse (R.6b)) der Entstehungszeitpunkt t_0 der zugehörigen Realweltausprägung gespeichert sein. Für den relevanten Fall eines unbekannten Entstehungszeitpunkts t_0 werden dagegen keine Handlungsempfehlungen gegeben.
Even/Shankaranarayanan (2007) stellen einen nutzenbasierten Ansatz zur Messung der Aktualität vor. Die Speicherdauer $t' \in R_0^+$ des betrachteten Attributwerts ω wird dabei als bekannt

angenommen und ist als Differenz zwischen dem Bewertungszeitpunkt t_1 und dessen Erfassungszeitpunkt t_0' definiert. Als Metrik wird eine Transformation der Speicherdauer t' auf das Intervall [0; 1] vorgeschlagen, die den aus der Aktualität des Attributwerts ω resultierenden, anwendungskontextspezifischen Nutzen ausdrücken soll. Die Autoren führen hierfür zwei verschiedene, alternativ zu sehende Nutzenfunktionen aus: Term (3) unterstellt, dass der Nutzen eines Attributwerts ω mit seiner Speicherdauer t' exponentiell abnimmt. Dabei stellt $\eta(A) \in R^+$ den zugehörigen Verfallparameter dar (je größer $\eta(A)$ ist, desto schneller nimmt der Nutzen mit zunehmender Speicherdauer t' ab).

$$Q^{\omega}_{Akt.}(t', A) = e^{-\eta(A) \cdot t'} \tag{3}$$

Daneben wird mit Term (4) eine alternative Nutzenfunktion vorgeschlagen, die auf der Annahme beruht, dass ein Attributwert ω bei Erreichen einer bekannten, festen maximalen Gültigkeitsdauer $T'_{\max}(A) \in R^+$ seine Aktualität und damit seinen Nutzen verliert. Die maximale Gültigkeitsdauer $T'_{\max}(A)$ wird hierbei zur Speicherdauer $t' \in R_0^+$ in Beziehung gesetzt. Über den Exponent $s \in R^+$ lässt sich zudem die Auswirkung des Quotienten aus t' und $T'_{\max}(A)$ auf das Metrikergebnis beeinflussen.

$$Q^{\omega}_{Akt.}(t', A) = \begin{cases} 1 - \left(\dfrac{t'}{T'_{\max}(A)} \right)^s & \text{für } 0 \le t' < T'_{\max}(A) \\ 0 & \text{sonst} \end{cases} \tag{4}$$

Für beide Funktionen sind die Ergebnisse auf einen beschränkten zusammenhängenden Wertebereich normiert (Infimum von null und Supremum von eins). Insofern sind die Metrikergebnisse für einen inaktuellen Attributwert und damit den minimalen Nutzen sowie für einen aktuellen Attributwert und damit den maximalen Nutzen eindeutig definiert (R.1). Bezüglich R.2 ist festzuhalten, dass die Metrikergebnisse nur im Fall einer kardinalen Nutzenfunktion auch selbst kardinalskaliert sind. Dieser Fall wird von Even/Shankaranarayanan (2007) jedoch nicht diskutiert. Vielmehr sprechen die Autoren von einem abstrakten Nutzen, sodass die Kardinalskalierung für die beiden Funktionen nicht eindeutig geprüft werden kann. Bei einem abstrakten Nutzen ohne Präferenzstärke fehlt jedoch die Dimension oder Maßeinheit im Sinne von R.3 (wie ist bspw. ein Nutzen von 0,5 zu interpretieren?). Dabei schließt die Normierung des Nutzens auf]0; 1] bzw. [0; 1] eine unmittelbare Interpretation in monetären Einheiten aus. Dagegen werden Vorschriften zur Aggregation der Metrikergebnisse auf Tupel-, Relationen- und Datenbankebene sowie für eine Menge von Datenbanken definiert, die eine konsistente Interpretation gewährleisten (R.4). Eine anwendungskontextspezifische Konfiguration der Metriken ist bei den exemplarischen Nutzenfunktionen durch entsprechende Wahl der Exponenten sowie der Aggregationsparameter möglich (R.5a)). Darüber hinaus wird angeführt, dass abhängig vom Anwendungskontext unterschiedliche Nutzenfunktionen für die Messung heranzuziehen sind, die bspw. auch wechselnde Verfallraten abbilden können. Insofern ist eine Konfiguration anhand attributspezifischer Charakteristika zwar möglich (R.5b)), ein Vorgehen zur Definition und Parametrisierung geeigneter (kardinaler) Nutzenfunktionen wird jedoch nicht vorgeschlagen. Eine Berücksichtigung von Zusatzdaten bei der Messung ist nicht vorgesehen (R.5c)). Bezogen auf die Ermittelbarkeit der Inputgrößen für die Metriken (R.6a)) ist festzuhalten, dass mit Ausnahme der *festen maximalen* Gültigkeitsdauer (vgl. Term (4)) die Inputgrößen grundsätzlich ermittelbar sind. Wie oben bereits diskutiert, verfügen viele Attribute jedoch über keine *feste maximale* Gültigkeitsdauer. Wenn die Inputgrößen allerdings ermittelt sind, ist die Berechnung der Metrikergebnisse automatisierbar (R.6b)). Tab. 3 gibt einen Überblick über die Ergebnisse der Analyse:

Anforderung	*Hinrichs (2002)*	*Ballou et al. (1998)*	*Even/Shankaranarayanan (2007)*
R.1 [*Normierung*]	Normierung auf einen beschränkten, zusammenhängenden Wertebereich	Normierung auf einen beschränkten, zusammenhängenden Wertebereich (durch Maximumfunktion)	Normierung auf einen beschränkten, zusammenhängenden Wertebereich
R.2 [*Kardinalskalierung*]	Nicht gewährleistet	Nur im Einzelfall bei Annahme einer gleichverteilten Gültigkeitsdauer und für s=1	Nur im Einzelfall bei Verwendung einer kardinalskalierten Nutzenfunktion
R.3 [*Interpretierbarkeit*]	Nicht gewährleistet	Nur im Einzelfall bei Annahme einer gleichverteilten Gültigkeitsdauer und für s=1	Nur im Einzelfall bei Verwendung einer kardinalskalierten Nutzenfunktion
R.4 [*Aggregierbarkeit*]	Mit Ausnahme der Datenbankebene konsistent möglich (d. h. auf Tupel- und Relationenebene)	Nicht gewährleistet	Auf Tupel-, Relationen- und Datenbankebene möglich
R.5 [*Konfigurierbarkeit*]			
a) [*Gewichtung*]	Anwendungskontext-spezifisch konfigurierbar durch Gewichtung bei der Aggregation (Ausnahme: Datenbankebene)	Anwendungskontext-spezifisch konfigurierbar mittels des Exponenten s	Anwendungskontext-spezifisch konfigurierbar mittels der Exponenten und durch Gewichtung bei der Aggregation
b) [*Attributspezifische Charakteristika*]	Nicht gewährleistet	Nicht gewährleistet	Bedingt realisierbar durch Wahl einer entsprechenden Nutzenfunktion
c) [*Berücksichtigung von Zusatzdaten*]	Nicht gewährleistet	Nicht gewährleistet	Nicht gewährleistet
R.6 [*Operationalisierbarkeit*]			
a) [*Ermittelbarkeit der Inputgrößen*]	Inputgrößen i. d. R. ermittelbar	Nur im Einzelfall, da viele Attribute über keine feste maximale Gültigkeitsdauer verfügen	Inputgrößen mit Ausnahme einer festen maximalen Gültigkeitsdauer ermittelbar
b) [*Automatisierbarkeit*]	Berechnung ist automatisierbar	Nur im Einzelfall, wenn der Entstehungszeitpunkt t_0 der zugehörigen Realweltausprägung gespeichert ist	Berechnung ist automatisierbar

Tab. 3 Zusammenfassung der Analyse bestehender Aktualitätsmetriken

Zusammenfassend legt die Analyse bestehender Ansätze offen, dass die Herausforderungen bei der Entwicklung einer neuen Metrik für Aktualität insbesondere in der Erfüllung der Anforderungen R.2, R.3 sowie R.5b) und R.5c) liegen. So sind die Kardinalskalierung und Interpretierbarkeit der Metrikergebnisse bisher, wenn überhaupt, nur in Einzelfällen gegeben. Sie stehen somit wegen ihrer jeweiligen Bedeutung (vgl. Abschnitt 3.1) im Fokus. Zudem erlaubt nur der Ansatz von Even/Shankaranarayanan (2007) eine Konfigurierbarkeit der Metrik im Sinne von Anforderung R.5b). Die Berücksichtigung von Zusatzdaten gemäß R.5c) wird von keiner bestehenden Metrik unterstützt.

4. Konstruktion einer neuen Metrik für Aktualität

Im Weiteren wird eine neue Metrik für Aktualität vorgestellt. Abschnitt 4.1 beschreibt zunächst die grundlegenden Überlegungen zur Konstruktion einer wahrscheinlichkeitsorientierten Metrik. Anschließend wird das Grundmodell der Metrik auf Attributwertebene entwickelt

(Abschnitt 4.2). Abschnitt 4.3 behandelt im Zuge einer Erweiterung des Grundmodells um den Fall fehlender Zusatzdaten einen weiteren relevanten Problembereich. In Abschnitt 4.4 werden zunächst Vorschriften zur Aggregation der Metrikergebnisse definiert, bevor in Abschnitt 4.5 verschiedene Möglichkeiten dargestellt werden, um die allgemeine Metrik für verschiedene Anwendungskontexte zu instanziieren.

4.1. Grundlegende Überlegungen zur Konstruktion der Metrik

Initial wurden verschiedene Alternativen analysiert, um die Normierung, Kardinalskalierung und Interpretierbarkeit der Metrikergebnisse zu gewährleisten. Ausgehend von Anforderung R.1, umschließt der Lösungsraum prinzipiell alle Funktionen, deren Wertebereich ein Infimum von null und ein Supremum von eins aufweist und die zumindest – wie auch die oben genannten existierenden Metriken – einen funktionalen Zusammenhang zwischen dem Alter des betrachteten Attributwerts und dem Metrikwert definieren (siehe Definition der Dimension Aktualität). Funktionsklassen, die beispielhaft hierunter fallen, sind parametrisierte Wurzelfunktionen ($f(t) = \left(1 + \sqrt{t}\right)^{-1}$ bzw. $f(t) = \left(\sqrt{1+t}\right)^{-1}$) oder der parametrisierte Arcustangens $\left(f(t) = 1 - \frac{2 \cdot ArcTan(t)}{\pi} \right)$, wobei $t \in R_0^+$ jeweils als Alter des betrachteten Attributwerts definiert ist. Zwar besitzen diese Funktionen einen beschränkten Wertebereich, allerdings werden andere Anforderungen wie die Kardinalskalierung und v. a. die Interpretierbarkeit resultierender Metrikergebnisse nicht erfüllt.

Als vielversprechend stellte sich demgegenüber eine Fundierung der Metrik auf wahrscheinlichkeits- und entscheidungstheoretischen Grundlagen heraus. Dies lässt sich wie folgt begründen: Aktualität ist die Eigenschaft eines Attributwerts ω, dass dieser zum Bewertungszeitpunkt noch den realen Gegebenheiten entspricht, wobei auf einen Realweltabgleich explizit verzichtet wird. Folglich liefert das Ergebnis einer Aktualitätsmetrik keine Aussage unter Sicherheit, sondern stellt vielmehr eine Schätzung dar. Vor diesem Hintergrund eignen sich grundsätzlich die Erkenntnisse der Wahrscheinlichkeitstheorie, welche die mathematischen Methoden zur Beschreibung und Untersuchung von Aussagen unter Unsicherheit liefert. Die grundlegende Idee besteht deshalb darin, Aktualität als Wahrscheinlichkeit dafür zu interpretieren, dass ein Attributwert ω zum Bewertungszeitpunkt t_1 noch der zugehörigen Realweltausprägung entspricht. Unter Voraussetzung einer begrenzten Gültigkeitsdauer des Attributwerts ω verringert sich so mit zunehmendem Alter t die Wahrscheinlichkeit (Metrikergebnis), dass der Wert ω noch aktuell ist.

Die Interpretation als Wahrscheinlichkeit hat wesentliche Vorteile: Erstens sind gerade für die Messung der Aktualität, die im Gegensatz zur Korrektheit einer Schätzung entspricht, eine Interpretation und Dimensionierung als Wahrscheinlichkeit sinnvoll (es ist keine Aussage unter Sicherheit möglich, da kein Realweltabgleich erfolgt). Zweitens können die Metrikergebnisse als Wahrscheinlichkeiten auch die Berechnung des Erwartungswerts von Entscheidungsalternativen und somit entscheidungstheoretische Verfahren unterstützen (vgl. auch Abschnitt 5.2). Drittens lassen sich Zusatzdaten methodisch fundiert berücksichtigen, indem das Metrikergebnis als bedingte Wahrscheinlichkeit (Zusatzdatum als Bedingung) dafür definiert wird, dass der betrachtete Attributwert noch den realen Gegebenheiten entspricht. Im Folgenden ist basierend auf diesen Überlegungen ein Grundmodell der Metrik für Aktualität dargestellt.

4.2. Grundmodell der Metrik auf Attributwertebene

Für das Grundmodell liegen folgende Annahmen und Definitionen zugrunde:

A.1 Ein Attributwert ω besitzt zunächst einen bekannten Entstehungszeitpunkt t_0 der zugehörigen Realweltausprägung sowie eine unbekannte und begrenzte[5] Gültigkeitsdau-

er $T \in R^+$, in welcher der Attributwert ω den realen Gegebenheiten entspricht. Die Gültigkeitsdauer T wird als zufällig angesehen (stetige Zufallsvariable). Die Messung der Aktualität des Werts ω erfolgt zum Zeitpunkt t_1 (mit $t_1 \geq t_0$).

Das Alter $t \in R_0^+$ des Attributwerts ω ergibt sich damit aus der Differenz zwischen dem Bewertungszeitpunkt t_1 und dem Entstehungszeitpunkt t_0. Ein Wert ω ist genau dann aktuell, wenn er zum Zeitpunkt t_1 noch den realen Gegebenheiten entspricht. Dies ist der Fall, wenn die Gültigkeitsdauer T größer oder gleich dem Alter t ist, wobei Alter und Gültigkeitsdauer in den gleichen Zeiteinheiten gemessen werden (bspw. in Jahren). Da die Gültigkeitsdauer T (in der Realität i. d. R.) unbekannt ist und deshalb als zufällig angesehen wird, kann die Aktualität des Attributwerts ω nicht unter Sicherheit ermittelt werden. Folglich wird im Weiteren unter Aktualität die Wahrscheinlichkeit verstanden, dass die Gültigkeitsdauer T größer oder gleich dem Alter t des Attributwerts ω ist.

A.2 Die Verteilungsfunktion $F^{\omega}(t|w_1,...,w_n) := P^{\omega}(T \leq t | W_1 = w_1,...,W_n = w_n)$ der Gültigkeitsdauer T des Attributwerts ω ist gegeben[6]. Sie hängt von den Zusatzdaten w_i (mit $i=1,\ldots,n$) ab. Diese Zusatzdaten w_i stellen Ausprägungen der Zufallsvariablen W_i dar und sind zunächst bekannt.

Auf Basis der Annahmen A.1 und A.2 ist die Aktualität eines Attributwerts ω als bedingte Wahrscheinlichkeit dafür definiert, dass dessen Gültigkeitsdauer T größer oder gleich dem Alter t ist. Die Zusatzdaten w_i werden dabei als Bedingung der Form $W_1 = w_1,...,W_n = w_n$ berücksichtigt. Die Ausprägungen der Variablen W_i sind zwar zunächst bekannt[7] und bräuchten nicht als Zufallsvariablen modelliert werden. Jedoch unterliegen diese auch einem zeitlichen Verfall. Deswegen ist es ohne Einschränkungen für das Grundmodell sogar vorteilhaft, diese im Hinblick auf spätere Modellerweiterungen als Zufallsvariablen zu modellieren. So kann z. B. die Wahrscheinlichkeit, dass der gespeicherte *Berufsstatus* „Student“ einer Person noch den realen Gegebenheiten entspricht, unter Einbeziehung von Zusatzdaten (bspw. der *Hochschulart* dieser Person) ermittelt werden. Die Metrik für Aktualität $Q^{\omega}_{Akt.}(t, w_1,..., w_n)$ ergibt sich damit wie folgt:

$$\begin{aligned} Q^{\omega}_{Akt.}(t, w_1,..., w_n) &:= P^{\omega}(T \geq t \mid W_1 = w_1,..., W_n = w_n) = 1 - P^{\omega}(T < t \mid W_1 = w_1,..., W_n = w_n) \\ &= 1 - F^{\omega}(t \mid w_1,..., w_n) = 1 - \int_0^t f^{\omega}(\theta \mid w_1,..., w_n) d\theta \end{aligned} \tag{5}$$

Das Metrikergebnis stellt allgemein die Wahrscheinlichkeit dafür dar, dass der Attributwert ω zum Bewertungszeitpunkt t_1 noch den realen Gegebenheiten entspricht. Dieser Wert wird über die Gegenwahrscheinlichkeit $P^{\omega}(T<t|W_1=w_1,...,W_n=w_n)$ (d. h. die Wahrscheinlichkeit, dass der Attributwert bis zum Betrachtungszeitpunkt t_1 bereits veraltet ist und $T<t=t_1-t_0$ gilt) sowie mithilfe der Verteilungsfunktion $F^{\omega}(t|w_1,...,w_n)$ ermittelt[8]. Die Verteilungsfunktion ist wiederum als Integral über die bedingte Wahrscheinlichkeitsdichtefunktion $f^{\omega}(\theta|w_1,...,w_n)$ definiert, die sich für $f^{\omega}(w_1,...,w_n)>0$ als Quotient aus den gemeinsamen Wahrscheinlichkeitsdichtefunktionen $f^{\omega}(\theta,w_1,...,w_n)$ und $f^{\omega}(w_1,...,w_n)$ ergibt. Weil sich die Gegenwahrscheinlichkeit darauf bezieht, dass der Attributwert ω inaktuell wird, bevor er das Alter t erreicht, ist das bestimmte Integral über den Bereich $[0; t]$ zu bilden.

Die neue Aktualitätsmetrik auf Attributwertebene kann somit auf Basis wahrscheinlichkeitstheoretischer Grundlagen definiert werden. Trotzdem unterliegt sie Restriktionen, da die Annahmen A.1 und A.2 die Realität lediglich partiell widerspiegeln. Zum einen ist hier die Annahme bekannter Entstehungszeitpunkte für die betrachteten Attributwerte (vgl. A.1) zu nennen. Oftmals wird bei der Erfassung eines Attributwerts ω in Unternehmen der Entstehungszeitpunkt t_0 der zugehörigen Realweltausprägung nicht gespeichert. In diesem Fall kann wie folgt verfahren werden: Datenbanken speichern i. d. R. den Erfassungszeitpunkt t_0' (mit $t_0 \leq t_0' \leq t_1$) des Attributwerts ω als Metadatum. Ist der Entstehungszeitpunkt t_0 unbekannt, kann demnach der bekannte Erfassungszeitpunkt t_0' zur Messung der Aktualität herangezogen werden[10,11]. Dies ist v. a. sinnvoll, wenn die Verteilung

$F'^{\omega}(t'|w_1,...,w_n) := P'^{\omega}(T' \leq t'|W_1=w_1,...,W_n=w_n)$ der Gültigkeitsdauer T' in Bezug auf den Zeitpunkt t_0' des Attributwerts ω ermittelbar ist (bspw. auf Basis historischer Daten). In diesem Fall ist die Wahrscheinlichkeit, dass der Attributwert ω noch den realen Gegebenheiten entspricht, in Abhängigkeit von seiner Speicherdauer t' ($t'=t_1-t_0'$) zu ermitteln. Dieser Wert stellt dann zugleich das Metrikergebnis für Aktualität $Q'^{\omega}_{Akt.}(t', w_1,...,w_n) := 1 - F'^{\omega}(t'| w_1,...,w_n)$ dar. Zum anderen wird in Annahme A.2 unterstellt, dass die relevanten Zusatzdaten für alle betrachteten Attributwerte gespeichert sind. Dies trifft in der Realität häufig nicht zu. Bspw. kann in einer Datenbank für einige Personen mit dem Attributwert „Student“ das Attribut *Hochschulart* bekannt sein, für andere wiederum nicht. Deshalb wird im Folgenden auch diese Annahme (vgl. A.2) relaxiert und das Grundmodell entsprechend erweitert.

4.3. Erweiterung des Grundmodells für den Fall fehlender Zusatzdaten

Um die Metrik für den Problembereich unbekannter Zusatzdaten w_i zu erweitern, wird A.2 wie folgt relaxiert:

A.2' Die Verteilungsfunktion $F^{\omega}(t|w_1,...,w_n) := P^{\omega}(T \leq t|W_1=w_1,...,W_n=w_n)$ der Gültigkeitsdauer T des Attributwerts ω ist gegeben. Sie hängt von den Zusatzdaten w_i (mit $i=1,\ldots,n$) ab. Diese Zusatzdaten w_i stellen Ausprägungen der Zufallsvariablen W_i dar, wobei zum Bewertungszeitpunkt t_1 lediglich $l \in IN$ ($l \leq n$) Zusatzdaten bekannt sind. Ohne Beschränkung der Allgemeinheit seien dies die Zusatzdaten w_j (mit $j=1,\ldots,l$).

Die Messung der Aktualität des Attributwerts ω erfolgt somit ohne Kenntnis der Ausprägungen der Zufallsvariablen W_k (mit $k=l+1,\ldots,n$). Da diese jedoch Teil der Verteilungsfunktion $F^{\omega}(t|w_1,...,w_n)$ (vgl. A.2') und damit der Metrik sind, entspricht diese Relaxierung des Grundmodells einer Messung der Aktualität, die mit weniger Zusatzdaten auskommen muss. In diesem Fall ist für die Zufallsvariablen W_k der Erwartungswert zu bilden, um die entsprechende Wahrscheinlichkeit bzw. das Metrikergebnis auch ohne Kenntnis der Ausprägungen w_k zu ermitteln.

So kann z. B. bei der Berechnung des Metrikergebnisses für den Attributwert „Student“, wenn die Ausprägung des Attributs *Hochschulart* für eine Person nicht bekannt ist, die über alle möglichen Ausprägungen der Zufallsvariable *Hochschulart* ermittelte Wahrscheinlichkeit Verwendung finden. Folglich geht nicht eine bestimmte Ausprägung des Attributs *Hochschulart*, wie z. B. *Universität,* in die Berechnung ein, sondern jede Ausprägung (also *Universität* und *Fachhochschule*) mit der zugehörigen Wahrscheinlichkeit. Für das oben genannte Beispiel würden die Wahrscheinlichkeiten herangezogen werden, dass der Student an einer *Universität* bzw. *Fachhochschule* studiert.

Durch die oben genannte Erwartungswertbildung ist auch weiterhin eine automatisierbare Berechnung der Metrikergebnisse ohne manuelle Eingriffe gewährleistet. Allerdings führt das Fehlen von Zusatzdaten dazu, dass sich die Güte des Metrikergebnisses verringert. Gemäß Annahme A.2' folgt die Gültigkeitsdauer T des Attributwerts ω der Verteilung $F^{\omega}(t|w_1,...,w_n)$. Demzufolge ist diese Verteilung zur Beurteilung der Güte des Metrikergebnisses im Falle fehlender Zusatzdaten heranzuziehen. Wird nämlich aufgrund fehlender Zusatzdaten auf die angepasste Verteilung (nach Erwartungswertbildung bezüglich der Zufallsvariablen W_k) zurückgegriffen, so führt dies i. d. R. zu veränderten Metrikergebnissen. Da die Zusatzdaten w_k (mit $k=l+1,\ldots,n$) unbekannt sind, kann die Verteilung $F^{\omega}(t|w_1,...,w_n)$ nicht direkt als Referenz zur Beurteilung der Güte der angepassten Metrik herangezogen werden. Stattdessen ist unter Berücksichtigung aller potenziellen Ausprägungen der Zufallsvariablen W_k (mit $k=l+1,\ldots,n$) zu bestimmen, wie weit das Metrikergebnis für den Fall fehlender Zusatzdaten im Erwartungswert bzw. maximal vom bisherigen Metrikergebnis abweicht. Diese Abweichung ist als erwarteter bzw. maximaler Fehler interpretierbar, der aufgrund der fehlenden Zusatzdaten bei der Ermittlung des Metrikergebnisses entsteht.

4.4. Aggregation der Metrikergebnisse

Die folgenden Ausführungen basieren auf dem weit verbreiteten relationalen Datenbankmodell und zielen auf eine Messung der Aktualität auf Attributwert-, Tupel-, Relationen- sowie Datenbankebene ab (R.4). Bei der Aggregation der Metrikergebnisse ist zugleich auch zwingend die anwendungskontextspezifische Bedeutung der einzelnen Attributwerte (Gewichtung gemäß R.5a)) und damit die Konfigurierbarkeit der Metrik zu berücksichtigen. Dabei wird die Metrik für Aktualität so konstruiert, dass die Metrik auf Ebene d+1 (z. B. Tupelebene) auf der Metrik auf Ebene d (z. B. Attributwertebene) aufsetzt. Dies zielt auf die konsistente Interpretation und Repräsentation der Metrikergebnisse ab.
Sei Γ ein Tupel mit den Attributwerten $\Gamma.A_1,\ldots,\Gamma.A_{|A|}$ für die Attribute $A_1,\ldots,A_{|A|}$. Des Weiteren sei die relative Bedeutung des Attributs A_i in Bezug auf den Anwendungskontext jeweils mit $g_i \in [0; 1]$ und $\sum_{i=1}^{|A|} g_i > 0$ gegeben. Dann ist die Metrik auf Tupelebene folgendermaßen definiert, wobei sich das Alter t, die Zusatzdaten $w_1,\ldots,w_n$ sowie deren Anzahl n jeweils auf den Attributwert $\Gamma.A_i$ beziehen[12]:

$$Q_{Akt.}(\Gamma) := \frac{\sum_{i=1}^{|A|} Q_{Akt.}^{\Gamma.A_i}(t, w_1, ..., w_n) g_i}{\sum_{i=1}^{|A|} g_i} \tag{6}$$

Die Aktualität einer nicht leeren Relation oder eines Views R folgt darauf basierend als arithmetisches Mittel der Metrikergebnisse für die enthaltenen Tupel Γ_j (j=1,…,|Γ|) aus R:

$$Q_{Akt.}(R) := \frac{\sum_{j=1}^{|\Gamma|} Q_{Akt.}(\Gamma_j)}{|\Gamma|} \tag{7}$$

Sei D eine Datenbank (oder eine Aggregation mehrerer Relationen oder Views), die sich als disjunkte Zerlegung der Relationen R_k (k=1,…,|R|) darstellen lässt (d. h., $D=R_1 \cup \ldots \cup R_{|R|}$ und $R_i \cap R_j = \emptyset \ \forall i \neq j$). Dann ist die Aktualität der Datenbank D auf Basis der Aktualität der Relationen R_k wie folgt definiert:

$$Q_{Akt.}(D) := \frac{\sum_{k=1}^{|R|} Q_{Akt.}(R_k) h_k}{\sum_{k=1}^{|R|} h_k} \tag{8}$$

Hierbei gewährleisten die Gewichte $h_k \in [0; 1]$ mit $\sum_{k=1}^{|R|} h_k > 0$ eine konsistente Interpretation und Repräsentation der Metrikergebnisse im Sinne von Anforderung R.4. Bspw. sind Relationen mit einer größeren Anzahl enthaltener Attribute und Tupel sowie einer größeren Bedeutung im Anwendungskontext stärker zu gewichten (vgl. R.5a)). Daneben resultiert eine konsistente, wahrscheinlichkeitsorientierte Interpretation als (gewichtete) durchschnittliche Wahrscheinlichkeit dafür, dass die zu bewertenden Attributwerte zum Bewertungszeitpunkt t_1 noch den realen Gegebenheiten entsprechen.

4.5. Möglichkeiten zur Instanziierung der Metrik

Im Folgenden werden unterschiedliche Möglichkeiten zur Instanziierung der Metrik dargestellt. Bei bekannten Zusatzdaten[13], sind die Verfallrate eines Attributs und dabei insbesondere die Verteilungsfunktion $F^{\omega}(t|w_1,\ldots,w_n) := P^{\omega}(T \leq t | W_1 = w_1, \ldots, W_n = w_n)$ zu bestimmen. Dafür werden vier generelle Möglichkeiten kurz diskutiert:

- Analyse öffentlich zugänglicher Daten (bspw. des Statistischen Bundesamts, anderer

öffentlicher Anstalten oder wissenschaftlicher Institutionen)
- Analyse unternehmensinterner Daten (z. B. im Data Warehouse)
- Durchführung einer Studie (bspw. Befragung einer Stichprobe an Kunden)
- Heranziehen von Expertenschätzungen (z. B. für die Verfallrate eines Attributs)

Die erste Möglichkeit betrifft die Nutzung öffentlich zugänglicher Daten. So können bspw. für die Messung der Aktualität von Attributen wie *Familienstand* oder *Adresse* empirische Daten (Eheschließungs-/Ehescheidungsraten bzw. Häufigkeit eines Wohnortwechsels) des Statistischen Bundesamts herangezogen werden. Auch für Attribute wie *Nachname*, bei denen die Nutzung öffentlich zugänglicher Daten zunächst nicht naheliegend erscheint, können mithilfe von z. B. Eheschließungs-/Ehescheidungsraten (sogar unter Berücksichtigung des Alters eines Kunden) Verfallraten ermittelt werden. Sind solche öffentlich zugänglichen Daten nicht verfügbar, sind unternehmensinterne, historische Daten zu analysieren. Dies ist insbesondere auch dann möglich und notwendig, wenn es sich um unternehmensspezifische Datenattribute handelt. So lässt sich im nachstehenden Beispiel des Mobilfunkanbieters die Verteilungsfunktion für das Attribut *Derzeitiger_Vertragstarif* ohne Weiteres auf Basis historischer Kundendaten im Data Warehouse ableiten. In diesem Zusammenhang kann für historisierte Attributwerte eine bitemporale Zeitstempelung erfolgen. Dabei werden bspw. die Erfassungs- und Updatezeitpunkte t_0' und - falls bekannt - der Entstehungszeitpunkt t_0 eines Attributwerts in der Realwelt jeweils als Metadatum festgehalten. Anschließend kann anhand dieser Zeitstempelung festgestellt werden, wann in der Vergangenheit Aktualisierungen eines Attributwerts durchgeführt wurden und wie lange dieser im Zeitverlauf den realen Gegebenheiten entsprach[14]. Ferner können Studien oder Expertenschätzungen durchgeführt werden, falls keine öffentlichen oder unternehmensinternen Daten verfügbar sind. Im Rahmen einer Studie kann bspw. eine Stichprobe an Kunden befragt werden, um auf Basis der Ergebnisse Verfallraten für die Metrikinstanziierung zu bestimmen. Wird hierbei die Messung der Aktualität von Adressdaten betrachtet, so ist für die Stichprobe auszuwerten, wann und wie häufig einzelne Kunden im Betrachtungshorizont ihren Wohnort wechselten. Auf diese Weise lassen sich nicht nur die durchschnittliche Dauer der Gültigkeit einer Kundenadresse, sondern auch deren durchschnittliche Verfallrate mithilfe des Quotienten 1/(mittlere Gültigkeitsdauer einer Adresse) sowie die zugehörige Verteilungsfunktion ermitteln. Eine andere Möglichkeit stellen Expertenschätzungen dar. So können bspw. zur Ermittlung der Verteilung für das Attribut *Derzeitiger_Vertragstarif* oder zu deren Plausibilisierung Vertriebsbeauftragte befragt werden.

5. Evaluation der Metrik für Aktualität

Die Evaluation der neuen Metrik erfolgt hinsichtlich einer ökonomischen, konstruktionstechnischen und erkenntnistheoretischen Perspektive (Cleven et al. 2009, Frank 2007). Die ökonomische Perspektive betrachtet dabei u. a. den praktischen Mehrwert, der durch den Einsatz eines Artefakts entsteht. Aspekte in diesem Kontext sind, dass das betrachtete Artefakt zu einer verbesserten Effizienz eines Geschäftsprozesses oder zu einer besseren Unterstützung bestimmter Entscheidungssituationen führt. Die konstruktionstechnische Perspektive zielt darauf ab, zu prüfen, inwiefern die Metrik zuvor definierten Anforderungen genügt. Schließlich fokussiert die erkenntnistheoretische Perspektive u. a. die kritische Reflektion menschlichen Urteilens und die Rekonstruktion des Erkenntnisfortschritts.
Um diesen Perspektiven Rechnung zu tragen, wird die Metrik zunächst den Anforderungen R.1 bis R.6 gegenübergestellt und diesbezüglich ihre Eignung diskutiert (konstruktionstechnische Perspektive). Dies stellt eine gängige Evaluationsmethode dar (Siau/Rossi 2011, S. 252). Da die bestehenden Ansätze ebenfalls mit Bezug zu den Anforderungen analysiert wurden, lässt sich hier auch ein Erkenntnisfortschritt aufzeigen (erkenntnistheoretische Perspektive). So soll festgestellt werden, inwieweit die zuvor identifizierte Forschungslücke geschlossen

wurde und welcher Erkenntnisfortschritt im Vergleich zu bestehenden Ansätzen vorliegt (Riege et al. 2009, S. 74). Um daneben auch den praktischen Mehrwert zu verdeutlichen, bedarf es zudem einer Realisierung der Metrik in der Realwelt (Riege et al. 2009, S. 75). Hier illustriert das Fallbeispiel eines großen Mobilfunkanbieters eine explizite Anwendung und speziell auch den praktischen Mehrwert der Metrik (ökonomische Perspektive).

5.1. Evaluation der Metrik gegen die Anforderungen R.1 bis R.6

Zu R.1: Die Definition der Metrik basierend auf wahrscheinlichkeitstheoretischen Grundlagen stellt sicher, dass die Metrikergebnisse auf das zusammenhängende Intervall [0; 1] in den reellen Zahlen beschränkt sind. Die Erfüllung von R.1 ergibt sich direkt anhand der Verteilungsfunktion $F^{\omega}(t|w_1,...,w_n)$. So ist eine Verteilungsfunktion $F(t)$ dadurch charakterisiert, dass sie monoton wachsend und rechtsstetig ist sowie $\lim_{t\to-\infty} F(t)=0$ und $\lim_{t\to\infty} F(t)=1$ gelten. Für die als $1\text{-}F^{\omega}(t|w_1,...,w_n)$ definierte Metrik folgt damit $Q^{\omega}_{Akt.}(t,w_1,...w_n)\in[0;1]$, wobei eindeutig definiert ist, dass das Infimum von null den Wert für inaktuelle und das Supremum von eins den für aktuelle Attributwerte repräsentieren.

Zu R.2: Die Kardinalskalierung der neuen Metrik folgt aus der entsprechenden Eigenschaft der Verteilungsfunktion $F^{\omega}(t|w_1,...,w_n)$. So ordnet eine Verteilungsfunktion $F(t)$ allgemein jeder Zahl t die Wahrscheinlichkeit $F(t)=P(T\leq t)$ zu. Dadurch, dass die Metrik als $1\text{-}F^{\omega}(t|w_1,...,w_n)$ definiert ist, stehen die Metrikergebnisse somit nicht nur in einer Rangordnung. Vielmehr ist ermittelbar, in welchem Ausmaß sich je zwei verschiedene Metrikergebnisse unterscheiden: Dieses Ausmaß ergibt sich als Differenz der Wahrscheinlichkeiten. Damit ist die Differenz zwischen Metrikergebnissen exakt bestimmbar und aussagekräftig.

Zu R.3: Die Interpretierbarkeit der Metrikergebnisse wird durch die wahrscheinlichkeitstheoretische Fundierung sichergestellt. Dabei ist das Metrikergebnis $Q^{\omega}_{Akt.}(t,w_1,...,w_n)$ als Wahrscheinlichkeit zu interpretieren und besitzt die Maßeinheit eins. Letzteres bedeutet: Die Wahrscheinlichkeit für die Aktualität eines beliebigen Attributwerts ω entspricht dem Grenzwert der relativen Häufigkeit aktueller Attributwerte bei einer theoretisch unendlich großen Menge an Attributwerten ω' des gleichen Attributs. Konkret stellt das Metrikergebnis $Q^{\omega}_{Akt.}(t,w_1,...,w_n)$ die Wahrscheinlichkeit dafür dar, dass der betrachtete Attributwert ω zum Bewertungszeitpunkt noch den realen Gegebenheiten entspricht. Demzufolge bedeutet ein Wert von eins respektive null, dass der Attributwert ω mit Sicherheit aktuell bzw. nicht mehr aktuell ist. Dazwischen nimmt der Wert der Metrik aufgrund der Monotonie der Verteilungsfunktion $F^{\omega}(t|w_1,...,w_n)$ ab. Zusammenfassend sind eine eindeutige Interpretierbarkeit gewährleistet und R.3 erfüllt.

Zu R.4: In Abschnitt 4.4 wurden Aggregationsvorschriften definiert, um eine konsistente Interpretation der Metrikergebnisse auf Tupel-, Relationen- sowie Datenbankebene zu gewährleisten. Zugleich ist deren konsistente Repräsentation sichergestellt, d. h., die Metrikergebnisse sind jeweils Resultat einer Abbildung auf das beschränkte, zusammenhängende Intervall [0; 1]. Somit ist eine konsistente Messung der Aktualität von Daten auch über die Attributwertebene hinaus möglich.

Zu R.5: Auf die kontextspezifische Konfigurierbarkeit der Metrik zielen drei Punkte ab:

a) Erstens können einzelne Attribute entsprechend ihrer Bedeutung im Anwendungskontext gewichtet werden (R.5a)). Dies ist auf Tupel- und Datenbankebene mittels der Gewichte g_i bzw. h_k möglich (vgl. Abschnitt 4.4).

b) Zweitens wurde die Metrik so konstruiert, dass spezifische Charakteristika unterschiedlicher Attribute (z. B. konstante, steigende, fallende oder wechselnde Verfallra-

ten) berücksichtigt werden können (R.5b)). Limitierende Annahmen bestehender Ansätze bezüglich der Verfallrate oder einer festen maximalen Gültigkeitsdauer von Attributwerten sind somit nicht notwendig.

c) Drittens unterstützt die Metrik die Berücksichtigung von Zusatzdaten (R.5c)). Hierzu wurde mit bedingten Wahrscheinlichkeiten gearbeitet (vgl. Abschnitt 4.2), um die Aktualität unter Einbeziehung relevanter Zusatzdaten methodisch fundiert zu messen. Die Zusatzdaten müssen jedoch nicht vollständig bekannt sein. Vielmehr stellt eine Erweiterung der Metrik deren Eignung auch dann sicher (vgl. Abschnitt 4.3), wenn Zusatzdaten für eine beliebige Anzahl betrachteter Attributwerte unbekannt sind.

Zu R.6: Für die Operationalisierbarkeit der Metrik sind folgende Aspekte sicherzustellen:

a) Zum einen müssen gemäß R.6a) alle Inputgrößen operationalisierbar und ermittelbar sein. Bezogen auf die neue Metrik sind zwei Inputgrößen zu diskutieren: die Verteilungsfunktion $F^{\omega}(t|w_1,...,w_n)$ sowie die Meta- und Zusatzdaten für die einzelnen Attributwerte. Für die Ermittlung der Verteilungsfunktion ist es notwendig, die attributspezifischen Verfallraten zu bestimmen. Welche Möglichkeiten hierfür existieren und wie dabei vorzugehen ist, wird in Abschnitt 4.5 dargestellt. Bei den Meta- und Zusatzdaten sind Entstehungszeitpunkt t_0 sowie Zusatzdaten w_i zu unterscheiden. Beide können, falls diese gespeichert sind, unkompliziert bspw. durch SQL DML Statements aus der Datenbank ausgelesen werden. Sollte der Entstehungszeitpunkt t_0 nicht gespeichert sein, ist der Erfassungszeitpunkt t_0' heranzuziehen. Dieser wird in Datenbanken gewöhnlich als Metadatum gespeichert. Ähnlich ist auch bei den Zusatzdaten zu verfahren: Sind diese für einzelne Attributwerte bekannt und gespeichert, so lassen sich diese unkompliziert aus der Datenbank auslesen. Sind die Zusatzdaten dagegen für beliebig viele Attributwerte nicht bekannt, ist auf die Metrikerweiterung in Abschnitt 4.3 zurückzugreifen. Allerdings gilt es jeweils zu bedenken, dass das Fehlen von Meta- bzw. Zusatzdaten, wie bereits diskutiert, mit einer Verringerung der Güte der Metrikergebnisse einhergeht.

b) Zum anderen ist gemäß R.6b) die automatisierbare Berechnung der Metrikergebnisse zu gewährleisten. Die neue Metrik ermöglicht dies u. a. mittels SQL DML Statements. Folgende Schritte sind durchzuführen:

   - Berechnung des Alters bzw. der Speicherdauer der betrachteten Attributwerte
   - Berechnung der Metrikergebnisse auf Attributwertebene
   - Aggregation der Metrikergebnisse für verschiedene Attributwerte
   - Nutzung der Metrikergebnisse bspw. in einem Erwartungswertkalkül

   Liegen die Inputgrößen der Metrik vor, lassen sich die obigen Schritte und damit die Metrikergebnisse automatisch berechnen. Allerdings ist kritisch anzumerken, dass die Gewichte der betrachteten Attribute im Falle der Aggregation manuell zu bestimmen sind. Dieser (primär einmalige) Aufwand ist jedoch im Vergleich zu alternativen Vorgehensweisen zu sehen, die i. d. R. wesentlich aufwendiger sind (bspw. Realweltabgleich) oder aufgrund limitierender Annahmen für viele Attribute keine adäquate Messung der Aktualität zulassen. Zudem ist die Erwartungswertbildung vorteilhaft, da im Fall nicht bekannter Zusatzdaten dennoch eine Berechnung der Metrikergebnisse ohne manuelle Eingriffe möglich ist.

5.2. Instanziierung und Einsatz der Metrik am Beispiel eines Mobilfunkanbieters

Dieser Abschnitt diskutiert die Instanziierung und Anwendung der Metrik im Fallbeispiel eines Mobilfunkanbieters. In Zusammenarbeit mit Fach- und Data Warehouse-Abteilung waren im Kampagnenmanagement Datenqualitätsprobleme identifiziert worden: Bei einer Befragung hatte sich herausgestellt, dass vielen Kunden aufgrund schlechter Datenqualität

ungeeignete Kampagnenangebote unterbreitet worden waren. Deshalb wurde die neue Metrik für Aktualität angewendet, um schon im Vorfeld einer Kampagne Probleme zu analysieren, wie etwa eine inadäquate Kundenselektion. Der Mobilfunkanbieter wählte dabei als Anwendungskontext eine bevorstehende Studentenkampagne aus[15]. Neben den Kundenkontaktdaten ist bei einer solchen Kampagne das Attribut *Berufsstatus* wichtig, das als Kriterium für die Kundenselektion dient. Zudem können aus Gründen der Preisdifferenzierung nur Kunden das Angebot annehmen, die tatsächlich noch Student sind. Folglich ist bei einem veralteten Wert des Attributs *Berufsstatus* keine Annahme des postalisch unterbreiteten Angebots möglich. Zweck der Kampagne ist es, Kunden dazu zu bewegen, in einen neuen Mobilfunktarif zu wechseln, der wegen seiner längeren Laufzeit und eines Mindestumsatzes zu Rentabilitätssteigerungen führt. Aus Vertraulichkeitsgründen wurden die im Beitrag verwendeten Daten und Zahlen anonymisiert, ohne die Vorgehensweise und die Ergebnisse zu verändern.
Im ersten Schritt wurde das bisherige Kampagnenvorgehen hinsichtlich der Aktualität des Selektionskriteriums *Berufsstatus* untersucht, um im zweiten Schritt dieses Vorgehen unter Berücksichtigung von Datenqualitätsaspekten zu verbessern.

5.2.1. Bisherige Vorgehensweise bei der Kundenselektion

Bisher wurden beim Mobilfunkanbieter drei Schritte durchgeführt, um die Zielkunden für eine Kampagne zu selektieren: Zunächst wurden alle Kunden ermittelt, die das definierte Selektionskriterium (*Berufsstatus*=„Student“) aufwiesen, wobei es sich in der betrachteten Kampagne um ca. 170.000 Kunden handelte. Hier lag die Annahme zugrunde, dass die selektierten Kunden auch tatsächlich noch Studenten waren. Anschließend wurde in Schritt zwei für jeden dieser Kunden das Umsatzvolumen des Vorjahres ermittelt. Grund war, dass sich bei Angebotsannahme von Kunden mit hohem Umsatzvolumen auch ein höherer Kampagnenertrag erzielen lässt. Vereinfachend war dieser Ertrag bei Vertragsabschluss mit 6% des bisherigen Umsatzvolumens zu berechnen. Der Aufwand (bspw. für die Kundenansprache) wurde dagegen nicht berücksichtigt, da dieser für jeden Kunden in gleicher Höhe anfällt und jedenfalls (für die fokussierten, 30% umsatzstärksten Kunden) kleiner ist als der Ertrag bei Vertragsabschluss. Somit war der Aufwand für die Kundenselektion nicht entscheidungsrelevant. Im dritten Schritt wurde das Angebot den 30% umsatzstärksten Kunden (ca. 51.000 Kunden) unterbreitet. Diese Beschränkung zielte u. a. auf den Ertrag, hatte aber v. a. Exklusivitätsgründe und sollte die Treue dieser „besten Kunden“ stärken.
Anhand dieses Vorgehens wurde in der Vergangenheit bei vergleichbaren Kampagnen eine durchschnittliche Erfolgsquote von circa 9% erzielt. Deshalb wurde in der ex ante Kalkulation für die Kampagne angenommen, dass ca. 4.590 Kunden (9% der 51.000 selektierten Kunden) das Angebot akzeptieren. Bei einem prozentualen Ertrag von 6% des Umsatzvolumens (bei Vertragsabschluss) ergaben sich bei einem durchschnittlichen Umsatz der selektierten 51.000 Kunden von ca. 1.470 EUR der Kampagnenertrag u pro Angebotsannahme zu ca. 88,20 EUR (1.470 EUR · 6%) und der Gesamtertrag U zu ca. 404.838 EUR (4.590 Kunden · 88,20 EUR). Dieser Gesamtertrag U lag deutlich über dem Aufwand der Kampagne.
Ein Ansatzpunkt zur Verbesserung der Kampagnendurchführung wurde insbesondere in der Datenqualität gesehen, da sich hier bei einer Kundenbefragung zum *Berufsstatus* entsprechende Probleme gezeigt hatten.

5.2.2. Angepasste Vorgehensweise bei der Kundenselektion

Um nunmehr Datenqualitätsaspekte einzubeziehen, wurde die Aktualität des Selektionskriteriums (*Berufsstatus*=„Student“) mittels der neuen Metrik gemessen und das Ergebnis bei der Kundenselektion berücksichtigt. Dazu wurde berechnet, wie wahrscheinlich es ist, dass ein Kunde mit gespeichertem *Berufsstatus*=„Student“ tatsächlich noch Student ist (Metrikergebnis). So wurde die Entscheidung unterstützt, inwiefern es sinnvoll ist, bspw. einen Kunden

mit einer sehr geringen Wahrscheinlichkeit überhaupt in der Kampagne anzuschreiben, da dieser im Fall eines inzwischen veralteten *Berufsstatus* das Angebot nicht mehr annehmen darf (und darüber im Zweifel sogar noch verärgert ist). Da es sich bei der potenziellen Zielgruppe um 170.000 Kunden handelte, war die Überprüfung der Korrektheit der Daten (Realweltabgleich durch Kontaktaufnahme mit jedem Kunden) nicht praktikabel. Deshalb wurde die bisherige Vorgehensweise wie folgt angepasst:
In Schritt eins wurden zunächst wiederum alle Kunden ermittelt, die das Selektionskriterium aufwiesen. Anschließend erfolgte die kundenindividuelle Messung der Aktualität des Selektionskriteriums (der nächste Abschnitt verdeutlicht die Instanziierung der Metrik). Hierbei wurden gespeicherte Meta- und Zusatzdaten verwendet. D. h., neben dem Alter t des Attributwerts wurden auch Zusatzdaten w_i wie bspw. die *Hochschulart* und die *Studienfachgruppe* einbezogen, sofern diese für den jeweiligen Kunden in der Datenbank gespeichert waren. Im dritten Schritt gingen diese Metrikergebnisse (Wahrscheinlichkeiten) in ein Erwartungswertkalkül ein. So wurde für jeden Kunden der Erwartungswert des Kampagnenertrags $E(u)$ automatisiert berechnet. Auf diese Weise wurde ein Kunde, der aufgrund seines hohen Umsatzvolumens bei Vertragsabschluss zwar einen hohen Kampagnenertrag u versprach, jedoch nur noch mit einer geringen Wahrscheinlichkeit Student war, in Beziehung zu anderen Kunden mit einem geringeren Umsatzvolumen aber einer höheren Wahrscheinlichkeit gesetzt. Dies erlaubte eine erwartungswertbasierte Entscheidungsunterstützung, welche Kunden in der Kampagne anzuschreiben waren. Die Berechnung des Erwartungswerts erfolgte mit:

$$E(u) = Q^{\omega}_{Akt.}(t, w_1, ..., w_n) \cdot u \tag{9}$$

Auf Basis der Erwartungswerte wurden in Schritt vier die 30% umsatzstärksten Kunden mit dem höchsten *erwarteten* Kampagnenergebnis selektiert.

5.2.3. Instanziierung der Metrik und Berechnung der Metrikergebnisse

Neben dem Selektionskriterium *Berufsstatus* umfasste das Datenschema auch den *Zeitpunkt der Immatrikulation* (t_0), die *Hochschulart* (W_1) sowie die *Studienfachgruppe* (W_2) eines Kunden als Attribute. Die Ausprägungen w_1 der *Hochschulart* richteten sich nach dem deutschen Hochschulsystem und umfassten die Werte „Universität“ und „Fachhochschule“. Die Ausprägungen w_2 der *Studienfachgruppe* waren aus der Menge „Wirtschafts- und Sozialwissenschaften“, „Ingenieurwissenschaften“, „Mathematik und Naturwissenschaften“, „Rechtswissenschaften“, „Agrar-, Forst- und Ernährungswissenschaften“, „Lehramt“, „Sprach- und Kulturwissenschaften“, „Kunst“ sowie „Medizin und Gesundheitswissenschaften“. Diese Daten wurden nach Möglichkeit bei der Erfassung abgefragt, waren jedoch nicht für alle Kunden gespeichert.
Um die Aktualität des Attributwerts „Student“ zu messen, wurden alle Fälle unterschieden, die dazu führen, dass dieser Attributwert seine Aktualität verliert. Hier waren erfolgreicher Studienabschluss und Studienabbruch zu unterscheiden. Für beide Fälle mussten die zugehörigen Wahrscheinlichkeitsverteilungen in Abhängigkeit vom jeweiligen Studiensemester bestimmt werden, wobei die Ausprägungen der Zufallsvariablen W_1 und W_2 als Zusatzdaten zu berücksichtigen waren (in Form von Bedingungen). Die bedingte Verteilung der Gültigkeitsdauer T des Attributwerts „Student“, gemessen in Semestern, war aus öffentlich zugänglichen Daten des Deutschen Statistischen Bundesamts sowie der Hochschul-Informations-System (HIS) GmbH abzuleiten (vgl. Heublein et al. 2003, 2008, Statistisches Bundesamt 2004, 2005, 2006, 2007). Die Daten bezogen sich jeweils auf Studierende des Bundesgebiets.

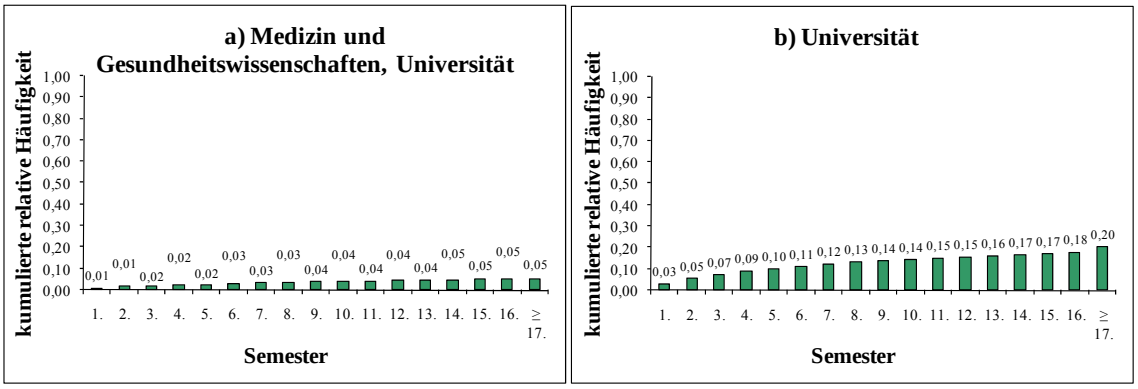

Abb. 4 Kumulierte relative Häufigkeiten von Studienabbrechern

Abb. 4 zeigt, bezogen auf die Studiendauer in Semestern, die kumulierten relativen Häufigkeiten für einen Studienabbruch an Universitäten exemplarisch für Studenten der Medizin und Gesundheitswissenschaften (links) sowie Studienfachgruppen übergreifend (rechts). Die kumulierten relativen Häufigkeiten in Abhängigkeit der *Hochschulart* und zusätzlich der *Studienfachgruppe* ließen sich dabei einfach für jede *Hochschulart* als Produkt aus dem Anteil der Studienabbrecher, die im jeweiligen Semester abbrechen, und der zugehörigen Studienabbruchquote innerhalb der *Studienfachgruppe* berechnen. Für die Instanziierung der Metrik im Grundmodell reichten diese Verteilungen bereits aus. Allerdings waren für den Fall fehlender Zusatzdaten auch Verteilungen zu ermitteln, die keine Kenntnis aller Zusatzdaten (z. B. *Studienfachgruppe*) voraussetzen. Deshalb wurden die kumulierten relativen Häufigkeiten eines Studienabbruchs auch für diesen Fall berechnet. Hierzu galt es, für die einzelnen *Hochschularten* den Erwartungswert hinsichtlich der unbekannten *Studienfachgruppe* zu bilden (vgl. Abschnitt 4.3). Für den hier betrachteten Fall wurde das gewichtete arithmetische Mittel der kumulierten relativen Häufigkeiten für die einzelnen *Studienfachgruppen* verwendet (diskreter Fall). Die Gewichtung richtete sich nach dem Anteil der Studierenden innerhalb der jeweiligen *Studienfachgruppe* bezogen auf die Gesamtheit aller Studierenden der betrachteten *Hochschulart*.

Mithilfe dieser Verteilungen war nun in Abhängigkeit der bekannten Zusatzdaten die Wahrscheinlichkeit zu berechnen, dass ein Kunde sein Studium nach einer Studiendauer von t Semestern bereits abgebrochen hat ($T \leq t$). Die Studiendauer t ergab sich dabei als Differenz zwischen dem Semester zum Bewertungszeitpunkt (t_1=Beginn Sommersemester 2009) und dem Zeitpunkt der Immatrikulation (t_0). Diese bedingte Wahrscheinlichkeit wird im Folgenden als Abbruchwahrscheinlichkeit $P_{Abbruch}^{"Student"}(T \leq t \mid W_1 = w_1,...,W_n = w_n)$ bezeichnet. Für einen Kunden A (vgl. Tab. 4), der an einer Universität vor elf Semestern (d. h. zum Wintersemester 2003/04 bei einem Bewertungszeitpunkt t_1=Beginn Sommersemester 2009) mit einem Studium der Medizin und Gesundheitswissenschaften begonnen hatte, betrug diese gemäß der oben ermittelten Verteilung bspw. nur 4% (vgl. Abb. 5 links).

Analog dazu wurden auch die Verteilungen für einen erfolgreichen Studienabschluss ermittelt. Abb. 5 zeigt für Universitäten wiederum exemplarisch die kumulierten relativen Häufigkeiten für Studenten der Medizin und Gesundheitswissenschaften (links) sowie Studienfachgruppen übergreifend (rechts) (Statistisches Bundesamt 2004, 2005, 2006, 2007): Zunächst wurden die Absolventenzahlen für die Berichtsjahre 2004 bis 2007 aggregiert, um zu einer umfangreichen Datengrundlage zu gelangen. Anschließend wurden aus den absoluten Häufigkeiten und dem jeweiligen Anteil der Absolventen in Bezug auf alle ursprünglich immatrikulierten Studenten die kumulierten relativen Häufigkeiten für einen Studienabschluss bedingt an die *Studienfachgruppe* sowie die *Hochschulart* errechnet (Heublein et al. 2008). Für den Fall fehlender Zusatzdaten waren diese Verteilungen entsprechend zu aggregieren (vgl. z. B. Abb. 5 rechts).

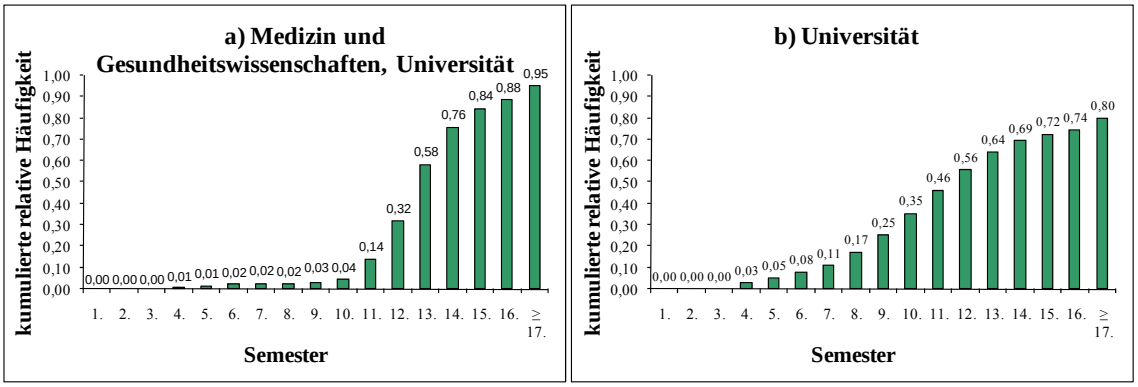

Abb. 5 Kumulierte relative Häufigkeiten von Studienabsolventen

Auf Basis der Verteilungen ließ sich in Abhängigkeit der bekannten Zusatzdaten die Wahrscheinlichkeit dafür bestimmen, dass ein Kunde nach *t* Semestern sein Studium bereits erfolgreich abgeschlossen hat. Diese wird mit $P^{"Student"}_{Abschluss}(T \le t \mid W_1 = w_1,...,W_n = w_n)$ bezeichnet und betrug für Kunde A nach elf Semestern zum Beispiel 14% (vgl. Abb. 5 links).
Abbruch- und Abschlusswahrscheinlichkeit waren nun wegen der sich ausschließenden Ereignisse zu addieren. Anschließend wurde die Gegenwahrscheinlichkeit gebildet, welche die Wahrscheinlichkeit repräsentiert, dass der Attributwert „Student“ noch aktuell ist. Die Aktualitätsmetrik ergab sich somit zu:

$$Q^{"Student"}_{Akt.}(t, w_1,..., w_n) := 1 - (P^{"Student"}_{Abbruch}(T \le t \mid W_1 = w_1,..., W_n = w_n) + P^{"Student"}_{Abschluss}(T \le t \mid W_1 = w_1,..., W_n = w_n)) \quad (10)$$

Für Kunde A errechnete sich das Metrikergebnis und damit die Wahrscheinlichkeit, dass dieser zum Betrachtungszeitpunkt, d. h. nach elf Semestern, tatsächlich noch studierte, zu 82%. Aber auch für Kunden, deren Meta- oder Zusatzdaten nicht vollständig erfasst waren, ließ sich das Metrikergebnis entsprechend ermitteln. Tab. 4 umfasst die Metrikergebnisse für ausgewählte Beispielkunden A bis D. Für Kunde D war das Metrikergebnis lediglich unter Berücksichtigung des Datums *Hochschulart*=„Universität“ zu bestimmen, da die Ausprägung der *Studienfachgruppe* nicht bekannt war (mit NULL gekennzeichnet). Hier ergab sich nach elf Semestern ein Metrikergebnis von 39%.

Kunde	Berufsstatus	Hochschulart	Studienfachgruppe	Semester *t*	$Q^{"Student"}_{Akt.}(t, w_1,..., w_n)$
A	**Student**	**Universität**	**Medizin und Gesundheitswissenschaften**	**11**	**82%**
B	**Student**	**Universität**	**Ingenieurwissenschaften**	**11**	**46%**
C	**Student**	**Fachhochschule**	**Ingenieurwissenschaften**	**11**	**18%**
D	**Student**	**Universität**	**NULL**	**11**	**39%**

Tab. 4 Beispieldaten für ausgewählte Kunden und zugehörige Metrikergebnisse

Im Vergleich: Angenommen es handelte sich bei Kunde D um einen Ingenieurwissenschaftler, so waren beim Metrikergebnis statt der 39% die korrekten 46% (vgl. Kunde B) auszuweisen. Wäre Kunde D für Medizin und Gesundheitswissenschaften immatrikuliert gewesen, so hätte sich die Aktualität zu 82% berechnet (vgl. Kunde A)[16]. Dies betont nochmals die Relevanz der Zusatzdaten w_i auf das Metrikergebnis und damit für die Messung der Aktualität.

5.2.4. Ergebnisse im Fallbeispiel

Mit der angepassten Vorgehensweise zur Kundenselektion ließ sich die Aktualität des Selektionskriteriums *Berufsstatus* berücksichtigen. Wie zu erwarten, waren die so selektierten 30% umsatzstärksten Kunden (51.000 Kunden) mit denen des bisherigen Vorgehens nicht deckungsgleich. Vielmehr ergab sich lediglich eine Schnittmenge von 20.130 Kunden, d. h. 30.870 Kunden wurden nur bei jeweils einem Vorgehen selektiert. Wegen dieser relativ großen Differenzmengen entschloss sich der Mobilfunkanbieter dazu, alle 81.870 Kunden anzu-

schreiben, die bei mindestens einer der beiden Vorgehensweisen ausgewählt wurden.
Die Ergebnisse der ex post Auswertung nach Durchführung der Kampagne waren folgendermaßen: Bei der Analyse der 51.000 Kunden, die mittels der bisherigen Vorgehensweise selektiert worden waren, ergab sich eine Erfolgsquote von lediglich ca. 7,7% (Vertragsabschluss mit 3.940 von 51.000 Kunden). Diese lag unter der erwarteten Erfolgsquote von 9%. Unter Verwendung des durchschnittlichen Umsatzes der 3.940 Kunden mit Vertragsabschluss von ca. 1.450 EUR, resultierte so ein ex post Kampagnenertrag E von ca. 342.780 EUR (=3.940 Kunden · 1.450 EUR · 6%).
Analog wurden auch die Kunden ausgewertet, die mittels der angepassten Vorgehensweise selektiert worden waren. Dabei ergab sich eine Erfolgsquote von ca. 17,3% (Vertragsabschluss mit 8.810 von 51.000 Kunden). Dies war per se nicht überraschend, da nun die Aktualität des Attributwerts „Student", die Voraussetzung für einen Vertragsabschluss ist, bei der Kundenselektion berücksichtigt worden war. Insofern nahmen hier insgesamt 8.810 Kunden das Angebot an. Obwohl hier der durchschnittliche Umsatz der Kunden mit Vertragsabschluss mit ca. 1.300 EUR erwartungsgemäß unterhalb dem des bisherigen Vorgehens lag, wurde insgesamt dennoch ein höherer Kampagnenertrag E von ca. 687.180 EUR (=8.810 Kunden · 1.300 EUR · 6%) erzielt. Folglich verdeutlicht die ex post Analyse, dass der Einsatz der Aktualitätsmetrik im Kampagnenmanagement zu besseren Ergebnissen führt als die bisherige Vorgehensweise.

5.2.5. Limitationen im Fallbeispiel

Die neue Metrik wurde zum einen hier nur für eine Kampagne des Mobilfunkanbieters angewendet. Deshalb ist das Vorgehen zukünftig auch in anderen Bereichen (bspw. bei anderen Kampagnen und für andere Datenattribute, wie Adressdaten) zu wiederholen, um die Resultate weiter zu evaluieren. Zum anderen wurde für die Kundenselektion unterstellt, dass der *Berufsstatus* „Student" bereits erfasst ist. Insofern blieben bisher Kunden unberücksichtigt, die nach dem Zeitpunkt der Erfassung in der Datenbank (bspw. mit dem *Berufsstatus* „Schüler") ein Studium begonnen haben. Diese Kunden sind als Zielgruppe der Kampagne ebenso interessant. Die Metrik ist daher auch für andere Werte des Attributs *Berufsstatus* anzuwenden. Dies ist ohne Weiteres möglich, da bspw. das Statistische Bundesamt Daten und Verteilungen zur Bestimmung der Wahrscheinlichkeiten für die Übergänge des *Berufsstatus* (bspw. von „Schüler" zu „Student") bereitstellt.

6. Zusammenfassung, kritische Würdigung und Ausblick

In der vorliegenden Arbeit wird eine wahrscheinlichkeitstheoretisch fundierte Metrik zur Messung der Aktualität von Datenwerten in Informationssystemen vorgeschlagen. Mithilfe dieser Metrik kann Aktualität zielgerichtet und weitgehend automatisiert gemessen werden, um ein ökonomisch orientiertes Datenqualitätsmanagement zu unterstützen. Zudem können allgemein die Ergebnisse der neuen Datenqualitätsmetrik erstmalig methodisch fundiert in Entscheidungen auf Grundlage von Erwartungswertkalkülen eingebunden werden. Die Operationalisierbarkeit der Metrik sowie ihren praktischen Mehrwert demonstriert eine Anwendung im Kampagnenmanagement eines Mobilfunkanbieters. Wesentliche Ergebnisse des Beitrags sind:

- Grundidee der neuen Metrik für Aktualität ist, die resultierenden Ergebnisse als Wahrscheinlichkeiten zu interpretieren. Deshalb kann eine methodisch fundierte Integration in Erwartungswertkalküle erfolgen. Zudem resultiert dadurch ein praktischer Mehrwert, der im Fallbeispiel anhand der Kenngrößen Erfolgsquote und Kampagnenertrag aufgezeigt wurde.
- Im Gegensatz zu bestehenden Ansätzen unterstützt die Metrik die Berücksichtigung rele-

vanter Zusatzdaten, die eine verbesserte Güte der Metrikergebnisse versprechen. Hier wird methodisch mit bedingten Wahrscheinlichkeiten gearbeitet. Zudem wurde, falls bei Datentupeln Zusatzdaten nicht bekannt sind, ein geeignetes Verfahren zur Erwartungswertbildung vorgestellt. Beides wurde im Fallbeispiel anhand der Zusatzdaten Hochschulart und Studienfachgruppe demonstriert.

- Der vorgeschlagene Ansatz kommt ohne limitierende Verteilungsannahmen und die Annahme einer festen, bekannten maximalen Gültigkeitsdauer aus. Damit ist die neue Metrik für eine Vielzahl von Attributen und deren Charakteristika einsetzbar.
- Durch die automatisierbare Berechnung der Metrikergebnisse wird der Ressourcenaufwand in Zusammenhang mit der Messung der Datenqualität gerade im Vergleich zu einem Realweltabgleich (vgl. Dimension Korrektheit) i. d. R. deutlich reduziert.

Mithilfe der vorgestellten Metrik ist es – im Vergleich zu existierenden Ansätzen – möglich, relevante Zusatzdaten zu berücksichtigen. Diese können jedoch eine schlechte Datenqualität (bspw. unvollständige Zusatzdaten) aufweisen, was kritisch zu diskutieren ist. Zwar können z. B. bei unvollständigen Zusatzdaten Erwartungswerte hinsichtlich der nicht bekannten Zusatzdaten gebildet werden. Dies führt aber – im Vergleich zu qualitätsgesicherten Zusatzdaten – zu tendenziell schlechteren Ergebnissen der Aktualitätsmessung. Eine weitere Limitation stellen die in Abschnitt 4.4 diskutierten Gewichte im Fall einer Aggregation der Metrikergebnisse dar. Diese sollen eine kontextspezifische Konfiguration ermöglichen, sind aber analog zu anderen Metriken manuell zu bestimmen. Dieser (primär einmalige) Aufwand der Definition der Gewichte ist jedoch im Vergleich zu alternativen Vorgehensweisen zu sehen, die i. d. R. wesentlich aufwendiger sind (bspw. Realweltabgleich) oder aufgrund von Limitationen für viele Attribute keine adäquate Aggregation zulassen. Auch die Ermittlung der Verteilungsfunktion als Basis für die Metrik (vgl. Abschnitt 4.5) gilt es kritisch zu diskutieren. Häufig kann hier zwar auf öffentlich zugängliche oder unternehmenseigene, historische Daten (z. B. des Data Warehouse) zurückgegriffen werden. Deren Auswertung ist unter Nutzung statistischer Datenanalyse-Software relativ schnell möglich. Aufwendiger ist die Metrikinstanziierung dagegen für Attribute, welche die explizite Durchführung einer Studie oder Expertenschätzungen erfordern. In diesem Fall gilt es abzuwägen, ob sich die Entwicklung einer Metrik unter ökonomischen Gesichtspunkten lohnt. Allerdings ist auch hier zu berücksichtigen, dass der Aufwand für die Ermittlung der Verteilungsfunktion dadurch relativiert wird, dass sowohl einmal ermittelte Inputgrößen als auch die entwickelte Metrikinstanz für mehrere Anwendungskontexte (bspw. bei Kampagnen, Kundenberatungen oder Produktentwicklungen) nutzbar sind.

Zudem gilt es zu bedenken, dass neben Aktualität noch weitere Dimensionen existieren, anhand derer Datenqualität zu messen ist. Auch wenn Aktualität als datenwertorientierte Dimension gerade aus fachlicher/betriebswirtschaftlicher Perspektive besonders relevant ist, so entspricht dies dennoch einer Partialsicht. Vor diesem Hintergrund sind zukünftig auch für andere Dimensionen wie Vollständigkeit, Korrektheit und Konsistenz geeignete Metriken zu konstruieren oder zu erweitern. Wegen ihrer Bedeutung ist hier die Dimension Vollständigkeit hervorzuheben (vgl. Al-Hakim 2007, S. 172, Klein/Callahan 2007, Lee et al. 2002, S. 134, Wand/Wang 1996, S. 96). Neben vorhandenen, aber inaktuellen Daten beeinflussen nämlich auch unvollständige Daten das Qualitätsniveau in einer Datenbank. Fehlen z. B. im Rahmen einer E-Mail-gestützten Marketingkampagne die E-Mail-Adressen einiger Kunden, so können diese nicht kontaktiert werden, auch wenn alle übrigen Kundendaten aktuell sind. Insofern können und müssen Metriken für derartige Qualitätsdimensionen zusammen mit der neuen Metrik für Aktualität als Unterstützung für ein ökonomisch orientiertes Datenqualitätsmanagement fungieren. Damit bilden sie einen notwendigen Schritt zur gezielten Auswahl ökonomisch sinnvoller Maßnahmen auf Basis des gemessenen Qualitätsniveaus. Mögliche Maßnahmen stellen Datenbereinigung oder der Zukauf von Adressdaten dar (vgl. hierzu im Detail z. B. Heinrich/Klier 2011). Wichtig dabei ist jedoch, dass sich derartige Qualitätsmetriken an den Anforderungen in Abschnitt 3.1 orientieren. Damit wäre eine notwendige

Basis für eine integrative Betrachtung der verschiedenen Qualitätsdimensionen gelegt.

Anmerkungen

[1] Die dargestellten kumulierten relativen Häufigkeitsverteilungen basieren auf Daten des Statistischen Bundesamtes und der Hochschul-Informations-System GmbH (HIS): (Heublein et al. 2003, 2008, Statistisches Bundesamt 2004, 2005, 2006, 2007).

[2] Zum Zweck der Illustration der Relevanz von Zusatzdaten werden im Beispiel lediglich diejenigen Studenten betrachtet, bei denen seit dem initialen Eintrag in einer (Unternehmens-)Datenbank zum Zeitpunkt der Immatrikulation keine Updates oder Überprüfungen erfolgt sind. Liegen abweichend davon bei anderen Studenten Updates oder Überprüfungen des Datenbankeintrags vor, kann die Messung der Aktualität natürlich ebenso erfolgen, allerdings dann unter Berücksichtigung dieser Zusatzdaten.

[3] Zusatzdaten sind nicht nur für die Dimension Aktualität, sondern auch für andere datenwertorientierte Dimensionen wie Vollständigkeit oder Konsistenz relevant.

[4] Der von Hinrichs (2002) verwendete Begriff „Update-Häufigkeit“ wurde hier aus Gründen der besseren Verständlichkeit durch den Begriff „Änderungshäufigkeit“ ersetzt.

[5] Bei unbegrenzter Gültigkeitsdauer T bleibt ein einmal erfasster Attributwert ω aktuell. Dieser Fall ist trivial, erfordert keine Messung der Aktualität und wird daher nicht betrachtet. Vielmehr wird eine Gültigkeitsdauer angenommen, die mit $T \in R^+$ zwar einen beliebig großen Wert annehmen kann, jedoch begrenzt ist. Im Gegensatz zu bestehenden Ansätzen (vgl. Abschnitt 3.2) ist die maximale Gültigkeitsdauer zudem nicht als fest und bekannt angenommen.

[6] Zur Ermittlung der Verteilungsfunktion vgl. Abschnitt 4.5.

[7] Diese Annahme ist auch dahingehend zu begründen, dass unter Wirtschaftlichkeitsaspekten zunächst bereits vorhandene Zusatzdaten (bspw. *Hochschulart* und/oder *Studienfachgruppe*) verwendet werden können.

[8] Die Gültigkeitsdauer T stellt gemäß A.1 eine stetige Zufallsvariable dar. Deshalb gilt $P^{\omega}(T<t|W_1=w_1,\ldots,W_n=w_n)=P^{\omega}(T\leq t|W_1=w_1,\ldots,W_n=w_n)=F^{\omega}(t|w_1,\ldots,w_n)$.

[9] Vgl. auch die Verwendung in den Ansätzen von Even/Shankaranarayanan (2007) und Hinrichs (2002).

[10] Unter der Annahme, dass der Attributwert zu diesem Zeitpunkt korrekt ist. Da die Daten bei der Erfassung meist vom Dateneigentümer angegeben werden, ist dies wenig restriktiv.

[11] Alternativ kann aufgrund bekannter Zusatzdaten auf t_0 und damit auf das Alter des Attributwerts t geschlossen werden. So ist es möglich, bspw. das Alter $t=t_1-t_0$ des Attributwerts „Student“ (Zeitdauer seit der Immatrikulation) auf Basis des zugehörigen Werts des Attributs *Geburtsdatum* zu schätzen (z. B. mithilfe historischer Daten).

[12] Auf eine zusätzliche Indizierung wurde aus Gründen der Übersichtlichkeit verzichtet.

[13] Die Berücksichtigung von Zusatzdaten ist dann sinnvoll, wenn diese für viele Datentupel bereits vorliegen. Eine gesonderte Erhebung ist dagegen nur selten zweckmäßig, da dann meist gleich die Attributwerte selbst erhoben werden könnten, für welche die Aktualität zu bestimmen ist.

[14] Für weitere Ausführungen zur Historisierung von Datenbeständen mittels temporaler Datenbanken sei an dieser Stelle auf Myrach (2005) verwiesen.

[15] Dies war auch deshalb naheliegend, da sich dieser Anwendungskontext bereits bei der ähnlichen Untersuchung von Heinrich et al. (2009) als geeignet herausgestellt hatte.

[16] Bei Betrachtung der Güte des Metrikergebnisses für den Kunden C ohne Zusatzdatum *Studienfachgruppe* ergab sich bezogen auf alle möglichen *Studienfachgruppen* ein erwarteter bzw. maximaler Fehler von immerhin etwa 12% bzw. 43%.

Literatur

Al-Hakim L (2007) Information quality factors affecting innovation process. Int J Inform Qual 1(2):162-176

Ballou DP, Pazer HL (1985) Modeling Data and Process Quality in Multi-Input, Multi-Output Information Systems. Manag Sci 31(2):150-162

Ballou DP, Tayi GK (1999) Enhancing Data Quality in Data Warehouse Environments. Comm ACM 42(1):73-78

Ballou DP, Wang RY, Pazer HL, Tayi GK (1998) Modeling Information Manufacturing Systems to Determine Information Product Quality. Manag Sci 44(4):462-484

Bamberg G, Baur F, Krapp M (2007) Statistik. Oldenbourg Wissenschaftsverlag, München

Batini C, Scannapieco M (2006) Data Quality. Concepts, Methodologies and Techniques (Data-Centric Systems and Applications). Springer, Berlin

Bertrand JWM, Fransoo JC (2002) Modelling and Simulation: Operations Management Research Methodologies using quantitative Modeling. Int J Oper Prod Manag 22(2):241-264

Bureau International des Poids et Mesures (2006) International System of Units (SI). Paris

Cappiello C, Francalanci C, Pernici B (2004) Data quality assessment from the user's perspective. In: Proceedings of the IQIS 2004, International Workshop on Information Quality in Information Systems, ACM, New York, 68-73

Cleven A, Gubler P, Hühner K (2009) Design Alternatives for the Evaluation of Design Science Research Artifacts. In: Proceedings of the 4th International Conference on Design Science Research in Information Systems and Technology, Malvern, PA

Cramer E, Kamps E (2008) Grundlagen der Wahrscheinlichkeitsrechnung und Statistik. Springer, Heidelberg

English LP (1999) Improving Data Warehouse and Business Information Quality: Methods for Reducing Costs and Increasing Profits. Wiley, New York

Eppler MJ (2003) Managing information quality. Springer, Berlin

Even A, Shankaranarayanan G (2007) Utility-Driven Assessment of Data Quality. Database Adv Inform Syst 38(2):75-93

Even A, Shankaranarayanan G, Berger PD (2007) Economics-Driven Data Management: An Application to the Design of Tabular Datasets. IEEE Trans Knowl Data Eng 19(6):818-831

Fahrmeir L, Künstler R, Pigeot I, Tutz G (2010) Statistik - Der Weg zur Datenanalyse. Springer, Heidelberg

Fisher CW, Chengalur-Smith IN, Ballou DP (2003) The Impact of Experience and Time on the Use of Data Quality Information in Decision Making. Inform Syst Res 14(2):170-188

Frank U (2007) Evaluation of Reference Models. In: Fettke P, Loos P (Hrsg) Reference Modelling for Business Systems Analysis. Idea Group Publishing, London

Franz T, von Mutius C (2008) Kundendatenqualität – Ein Schlüssel zum Erfolg im Kundendialog. In: Swiss CRM Forum 2008, Zürich

Grosser T, Bange C (2009) Datenqualität in SAP-Systemen - Eine unabhängige Anwenderbefragung über die Wahrnehmung der Datenqualität in SAP-Systemen. BARC-Institut, Würzburg

Gustavsson M, Wänström C (2009) Assessing information quality in manufacturing planning and control processes. Int J of Qual Reliab Manag 26(4):325-340

Harris Interactive (2006) Information Workers Beware: Your Business Data Can't Be Trusted. http://www.sap.com/about/newsroom/businessobjects/20060625_005028.epx. Abruf am 28 Feb 2011

Heinrich B, Klier M (2006) Ein Optimierungsansatz für ein fortlaufendes Datenqualitätsmanagement und seine praktische Anwendung bei Kundenkampagnen. Z Betriebswirtsch 76(6):559-587

Heinrich B, Klier M (2009) A Novel Data Quality Metric for Timeliness considering Supplemental Data. In: Proceedings of the 17th European Conference on Information Systems, Verona, 2701-2713

Heinrich B, Klier M (2011) Assessing Data Currency - A Probabilistic Approach. J Inform Sci 37(1):86-100

Heinrich B, Kaiser M, Klier M (2008) Does the EU Insurance Mediation Directive help to improve Data Quality? - A metric based analysis. In: Proceedings of the 16th European Conference on Information Systems, Gelway

Heinrich B, Kaiser M, Klier M (2009) A Procedure to Develop Metrics for Currency and its Application in CRM. ACM J Data Inf Quality 1(1):5:1-5:28

Helfert M (2002) Proaktives Datenqualitätsmanagement in Data-Warehouse-Systemen - Qualitätsplanung und Qualitätslenkung. Logos, Berlin

Heublein, U, Spangenberg, H, Sommer, D (2003) Ursachen des Studienabbruchs. http://www.bmbf.de/pub/ursachen_des_studienabbruchs.pdf. Abruf am 31 Jan 2011

Heublein, U, Schmelzer, R, Sommer, D (2008) Die Entwicklung der Studienabbruchquote an den deutschen Hochschulen: Ergebnisse einer Berechnung des Studienabbruchs auf der Basis des Absolventenjahrgangs 2006. http://www.his.de/pdf/21/his-projektbericht-studienabbruch_2.pdf. Abruf am 31 Jan 2011

Hinrichs H (2002) Datenqualitätsmanagement in Data Warehouse-Systemen. Dissertation Universität Oldenburg, Oldenburg

Jiang Z, Sarkar S, De P, Dey D (2007) A Framework for Reconciling Attribute Values from Multiple Data Sources. Manag Sci 53(12):1946

Juran JM (1998) How to think about Quality. In: Juran JM, Godfrey AB (Hrsg) Juran's Quality Handbook. McGraw-Hill, New York

Kaplan D, Krishnan R, Padman R, Peters J (1998) Assessing Data Quality in Accounting Information Systems. Comm ACM 41(2):72-78

Kengpol A (2006) Using Information Quality Techniques to Improve Production Planning and Control. Int J Manag 23(1):53-60

Klein BD, Callahan TJ (2007) A comparison of information technology professionals' and data consumers' perceptions of the importance of the dimensions of information quality. Int J Inform Qual 1(4):392-411

Kraus C (2004) Adress- und Kundendatenbanken für das Direktmarketing. Aufbau, Pflege, Nutzung. Businessvillage, Göttingen

Lee YW, Strong DM, Kahn BK, Wang RY (2002) AIMQ: a methodology for information quality assessment. Inform Manag 40(2):133-146

Lindner H, Siebke W, Simon G (2006) Physik für Ingenieure. Hanser Verlag, Leipzig

Meredith JR, Raturi A, Amoako-Gyampah K, Kaplan B (1989) Alternative Research Paradigms in Operations. J Oper Manag 8(4):297-326

Monczka RM, Petersen KJ, Handfield RB, Ragatz GL (1998) Success Factors in Strategic Supplier Alliances: The Buying Company Perspective. Decision Sciences 29(3):553-577

Myrach T (2005) Temporale Datenbanken in betrieblichen Informationssystemen: Prinzipien, Konzepte, Umsetzung. Teubner, Wiesbaden

Nicholas D, Herman E (2009) Assessing Information Needs: Tools, Techniques and Concepts for the Internet Age. Routledge, London

Orr K (1998) Data Quality and Systems Theory. Comm ACM 41(2):66-71

Parssian A (2006) Managerial decision support with knowledge of accuracy and completeness of the relational aggregate functions. Decis Support Syst 42(3):1494-1502

Parssian A, Sarkar S, Jacob VS (2004) Assessing Data Quality for Information Products: Impact of Selection, Projection, and Cartesian Product. Manag Sci 50(7):967-982

Parssian A, Sarkar S, Jacob VS (2009) Impact of the Union and Difference Operations on the Quality of Information Products. Inform Syst Res 20(1):99-120

Pipino L, Lee YW, Wang RY (2002) Data Quality Assessment. Comm ACM 45(4):211-218

Redman TC (1996) Data Quality for the Information Age. Artech House, Boston

Riege C, Saat J, Bucher T (2009) Systematisierung von Evaluationsmethoden in der gestaltungsorientierten Wirtschaftsinformatik. In: Becker J, Krcmar H, Niehaves B (Hrsg) Wissenschaftstheorie und gestaltungsorientierte Wirtschaftsinformatik. Physica-Verlag, Heidelberg

Russom P (2006) Taking Data Quality to the Enterprise through Data Governance. TDWI Report Series March 2006, The Data Warehousing Institute, Seattle

SAS Institute (2003) Europäische Unternehmen leiden unter Profitabilitätseinbußen und niedriger Kundenzufriedenheit durch schlechte Datenqualität. Studie der SAS Institute GmbH, Heidelberg

Schönfeld A (2007) ADS AdressDrehScheibe – regelbasierter Datenaustausch mit Open Source. In: Open Source Meets Business 2007, Nürnberg

Shankaranarayanan G, Cai Y (2006) Supporting data quality management in decision-making. Decis Support Syst 42(1):302-317

Siau K, Rossi M (2011) Evaluation techniques for systems analysis and design modelling methods - a review and comparative analysis. Info Systems J 21(3):249-268

Statistisches Bundesamt (2004-2007) Prüfungen an Hochschulen - Fachserie 11 Reihe 4.2 - 2003-2006. http://www.destatis.de. Abruf am 31 Jan 2011
Von Alven WH (1964) Reliability engineering. Prentice-Hall, Englewood Cliffs, NJ
Wand Y, Wang RY (1996) Anchoring data quality dimensions in ontological foundations. Comm ACM 39(11):86-95
Wang RY, Strong DM (1996) Beyond accuracy: what data quality means to data consumers. J Manag Inform Syst 12(4):5-33
Wang RY, Storey VC, Firth CP (1995) A Framework for analysis of data quality research. IEEE Trans Knowl Data Eng 7(4):623-640
West M (2011) Developing high quality data models. Morgan Kaufmann, Burlington

Data quality assessment – a metric-based approach to quantify the currency of data in information systems

Abstract: Due to the importance of using up-to-date data in information systems, this paper analyzes how the data quality dimension currency can be measured. Therefore, we design a probability based metric that allows for an objective and to a great extent automated assessment of data's currency. In contrast to existing approaches, the resulting values of the new metric meet important requirements such as ratio scale and can be interpreted as probabilities. Hence, they can also be applied to calculate expected values for decision making in a methodically well-founded manner. Moreover, the metric can be adapted to the context of a particular application considering both, the specific characteristics of attribute values and supplemental data stored in the information system. The evaluation of the approach is based on six requirements for data quality metrics. Furthermore, the case of a mobile services provider illustrates the metric's applicability and its practical benefit.

Keywords: Data quality, Currency, Metric, Assessment