Universität Augsburg
Prof. Dr. Hans Ulrich Buhl
Kernkompetenzzentrum
Finanz- & Informationsmanagement
Lehrstuhl für BWL, Wirtschaftsinformatik,
Informations- & Finanzmanagement

Universität Augsburg
University

Diskussionspapier WI-200

Metriken zur Bewertung der Datenqualität - Konzeption und praktischer Nutzen

von

Mathias Klier

in: Informatik-Spektrum 31 (2008) 3, S. 223-236

Universität Augsburg, 86135 Augsburg
Besucher: Universitätsstr. 12, 86159 Augsburg
Telefon: +49 821 598-4801 (Fax: -4899)
www.fim-online.eu

Elitenetzwerk Bayern
Universität Augsburg University
TECHNISCHE UNIVERSITÄT MÜNCHEN
BACHELOR WIN WIRTSCHAFTS INFORMATIK
ZEITSCHRIFTEN WIRTSCHAFTS INFORMATIK BUSINESS & INFORMATION SYSTEMS ENGINEERING

Metriken zur Bewertung der Datenqualität

Konzeption und praktischer Nutzen

Dipl.-Math. oec. Mathias Klier
Lehrstuhl für Betriebswirtschaftslehre, Wirtschaftsinformatik & Financial Engineering der Universität Augsburg, Universitätsstraße 16, 86159 Augsburg
Mathias.Klier@wiwi.uni-augsburg.de

[Zusammenfassung]

Seit einigen Jahren gewinnt das Thema Datenqualität (DQ) sowohl in der Wissenschaft als auch in der Praxis immer mehr an Bedeutung. Dies überrascht nicht, da die Sicherstellung und Verbesserung der DQ – insbesondere im Zuge von stark wachsenden Datenbeständen, dem verstärkten Einsatz von Data Warehouse-Systemen und der Relevanz, die dem Customer Relationship Management (CRM) beigemessen wird – eine immer größere Rolle spielt. Mit der steigenden Bedeutung der DQ wird auch die Notwendigkeit geeigneter Mess- und Bewertungsverfahren deutlich, die für eine Planung, Steuerung und Kontrolle von DQ-Maßnahmen unabdingbar sind. Um die DQ zu quantifizieren, werden im Beitrag neue Metriken für die DQ-Merkmale Korrektheit und Aktualität vorgestellt und diskutiert. Dabei wird im Vergleich zu bestehenden Ansätzen insbesondere Wert auf fachliche Interpretierbarkeit und Praxistauglichkeit gelegt. Die Anwendung der entwickelten Metriken und damit auch die Eignung im praktischen Einsatz werden im CRM-Kontext anhand des Kampagnenmanagements eines großen deutschen Mobilfunkanbieters veranschaulicht.

[Abstract]

In recent years data quality (DQ) has gained more and more importance in theory and practice due to an extended use of data warehouse systems, management information systems and a higher relevance of customer relationship management. This refers to the fact that for decision makers the benefit of data heavily depends on completeness, correctness and timeliness for example. The growing relevance of DQ revealed the need for adequate measurement because quantifying DQ (e. g. of a data base) is essential for planning DQ measures in an economic manner. The article analyzed how DQ criteria can be quantified in a goal-oriented and economic manner. The aim was to develop new metrics for the DQ criteria correctness and timeliness. These metrics proposed enable an objective and automated measurement. In contrast to existing approaches the metrics were designed according to important requirements like feasibility and interpretability. In cooperation with a major German mobile services provider, the developed metrics were applied and they turned out to be appropriate for practical problems.

[Vorspann]

„What doesn't get measured doesn't get managed" – natürlich trifft dies auch auf den Bereich der Datenqualität zu. Daher werden im Beitrag neue Metriken für die Datenqualitätsmerkmale Korrektheit und Aktualität entwickelt und deren Anwendung im Rahmen des Customer Relationship Managements eines Mobilfunkanbieters beschrieben.

Einleitung

In den vergangenen Jahren hat – insbesondere im Zuge des verstärkten Einsatzes von Data Warehouse-Systemen bspw. im Bereich des Customer Relationship Managements – Datenqualität (DQ) sowohl in der Wissenschaft als auch in der Praxis immer mehr an Bedeutung gewonnen. Die zunehmende Relevanz, die der Thematik beigemessen wird, überrascht dabei nicht, da der Nutzen der Versorgung von Entscheidungsträgern mit Daten mit deren Vollständigkeit, Korrektheit und Aktualität steigt bzw. fällt – also mit Eigenschaften, die als Qualitätskriterien bekannt sind [27]. Für viele Unternehmungen stellt dabei die Sicherstellung der DQ immer noch ein Problem dar [26], obwohl die „Total Cost of poor Data Quality“ laut Untersuchungen von Redman in einer Größenordnung zwischen 8 und 12 Prozent des Unternehmensumsatzes liegen [24]. Andere Zahlen besagen, dass sich bei Data Warehouse-Projekten wegen inkorrekter und fehlender Daten der Anteil am Budget für (geplante und vor allem ungeplante) DQ-Maßnahmen auf mehr als 50 Prozent beläuft [19, 1]. Die Auswirkungen einer schlechten DQ sind dabei vielfältig: Sie reichen von einer Verschlechterung der Kundenbeziehung und -zufriedenheit durch falsche Kundenansprache bis hin zu einer mangelhaften Entscheidungsunterstützung des Managements. All dies verdeutlicht, welche Bedeutung dem DQ-Thema vor allem in IT-Projekten heute zukommt.
Damit wird aber auch die Notwendigkeit geeigneter Mess- und Bewertungsverfahren deutlich. So weist Naumann explizit darauf hin, dass für die Planung von DQ-Maßnahmen unter Kosten-Nutzen-Gesichtspunkten Metriken für den jeweilig aktuellen Stand der DQ (bspw. bezogen auf eine Datenauswertung) unverzichtbar sind [20] (vgl. auch [13, 22]). Daher wird im Folgenden die Fragestellung aufgegriffen, wie Metriken für ausgewählte DQ-Merkmale entwickelt werden können. Diese sollen die Messung der DQ zum jeweiligen Analysezeitpunkt ermöglichen und die Untersuchung zukünftiger Auswirkungen auf das DQ-Niveau, wie z. B. bei der Durchführung von DQ-Maßnahmen (Data Cleansing etc.) oder den zeitlichen Verfall einzelner Datenwerte (bspw. Kundenadresse) unterstützen.
Der Beitrag ist folgendermaßen strukturiert: Im zweiten Kapitel werden Anforderungen an DQ-Metriken definiert, bevor im darauf folgenden Abschnitt auf bisherige Beiträge zur DQ-Messung eingegangen wird. Im vierten Kapitel werden eigene Metriken für die DQ-Merkmale Korrektheit und Aktualität entwickelt sowie deren Vorteile herausgearbeitet. Eine Beschreibung der praktischen Anwendung der Metrik für Aktualität im Rahmen des Kampagnenmanagements eines großen Mobilfunkanbieters findet sich im fünften Abschnitt, bevor im letzten Teil die Ergebnisse zusammengefasst werden.

Anforderungen an Datenqualitätsmetriken

Für ein ökonomisch orientiertes DQ-Management sind Metriken erforderlich, die eine Beurteilung von DQ-Maßnahmen unter Kosten-Nutzen-Gesichtspunkten ermöglichen. Die Zusammenhänge lassen sich anhand des Regelkreises in Abb. 1 veranschaulichen.

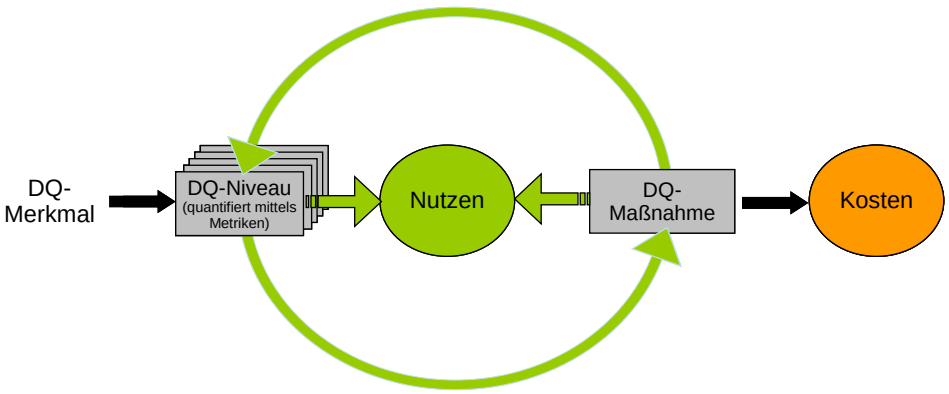

Abb. 1 Datenqualitätsregelkreis

Den Regler, über den in den Regelkreis eingegriffen werden kann, stellen die DQ-Maßnahmen dar. Die Durchführung von Maßnahmen hat dabei eine Verbesserung des DQ-Niveaus (zu quantifizieren mittels Metriken) zur Folge, womit ein entsprechender ökonomischer Nutzen, bspw. eine effektivere Kundenansprache im Rahmen einer Kampagne (vgl. hierzu z. B. das Anwendungskapitel dieses Beitrags), verbunden ist. Umgekehrt kann anhand des DQ-Niveaus und unter Einbeziehung von entsprechenden Richt- und Schwellenwerten über die Durchführung von (weiteren) Maßnahmen entschieden werden. Aus einer ökonomischen Betrachtungsweise heraus muss die Auswahl der Maßnahmen dabei unter Berücksichtigung von Kosten-Nutzen-Gesichtspunkten

erfolgen (vgl. z. B. [5, 10, 18, 25]) – bspw. würde man so bei zwei vorhandenen Maßnahmen, aus denen (annähernd) der gleiche ökonomische Nutzen resultiert, nur die kostengünstigere in Betracht ziehen.
Die Identifikation und Klassifikation von DQ-Merkmalen wird in einer Vielzahl von sowohl theoretisch- als auch anwendungsorientierten Veröffentlichungen thematisiert [4, 21, 28, 14, 17, 8, 16, 7, 23, 15]. Im Folgenden werden die Merkmale Korrektheit und Aktualität fokussiert, da sich hier noch keine Metriken etabliert haben und bspw. im Gegensatz zum DQ-Merkmal Vollständigkeit nur sehr wenige Ansätze und Ideen zu deren Quantifizierung existieren. Um eine theoretische Fundierung zu gewährleisten und eine praktische Anwendung zu ermöglichen, werden folgende Anforderungen an DQ-Metriken definiert (in Teilen ähnliche Anforderungen finden sich auch bei Even und Shankaranarayan [9], Hinrichs [15] und Heinrich et al. [12]):

F1: [*Normierung*] Um die Interpretierbarkeit und Vergleichbarkeit der Metrikergebnisse (bspw. bei Messung zu verschiedenen Zeitpunkten) zu gewährleisten, ist eine geeignete Normierung der Metrikergebnisse (bspw. auf das Intervall [0;1]) zu fordern.

F2: [*Kardinale Skalierung*] Um eine Betrachtung der zeitlichen Entwicklung der Metrikergebnisse und eine ökonomische Beurteilung von Maßnahmen zu unterstützen, ist es erforderlich, dass diese kardinal skaliert sind. Eine Kardinalskala im Sinne der Statistik liegt dabei dann 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“ [3][1].

F3: [*Sensibilisierbarkeit*] Um das DQ-Niveau zielgerichtet messen und beurteilen zu können, ist es notwendig, dass die Metriken für die konkrete Anwendung sensibilisiert und für die jeweilige Zielsetzung, welche der Messung zugrunde liegt, konfiguriert werden können (z. B. um wichtige Attribute wie die Telefonnummer des Kunden – im Falle einer Telefonkampagne – stärker gewichten zu können).

F4: [*Aggregierbarkeit*] Um bei Zugrundelegung eines relationalen Datenbankschemas einen flexiblen Einsatz zu ermöglichen, soll die Messung des DQ-Niveaus auf Attributwert-, Tupel-, Relationen- sowie Datenbankebene möglich sein – dabei soll zudem die Aggregierbarkeit der Metrikergebnisse auf einer Ebene zur nächst höheren Ebene gewährleistet sein. Demzufolge soll sich bspw. die Bewertung der Korrektheit einer Relation aus der Bewertung der Korrektheit der enthaltenen Tupel ergeben.

F5: [*Operationalisierbarkeit mittels Messverfahren*] Um die praktische Anwendung der Metriken zu ermöglichen, müssen ausgehend von den Metriken geeignete (z. B. bezüglich Definitionsbereich, Wertebereich etc.) Messverfahren definiert werden, welche die Metriken operationalisieren.

F6: [*Fachliche Interpretierbarkeit*] In der praktischen Anwendung reicht i. d. R. die bloße Normierung und Kardinalität der Metriken nicht aus. Vielmehr müssen die resultierenden Metrikergebnisse auch *fachlich* interpretierbar (bspw. als Anteil der korrekt erfassten Attributwerte in einer Datenbank) und reproduzierbar sein (bspw. sollten die Metrikergebnisse auch von Dritten nachvollzogen werden können).

Im folgenden Abschnitt werden basierend auf den obigen Anforderungen ausgewählte Ansätze zur Quantifizierung der DQ im Allgemeinen sowie der DQ-Merkmale Korrektheit und Aktualität im Speziellen vorgestellt.

Bisherige Beiträge zur Datenqualitätsmessung

In der Literatur findet sich bereits eine ganze Reihe von Ansätzen zur Bewertung des DQ-Niveaus, die sich neben den berücksichtigten Merkmalen vor allem in den zugrunde liegenden Messverfahren unterscheiden (für einen Überblick vgl. [27]).
Am Massachusetts Institute of Technology wurde die AIM Quality-Methode entwickelt [17]. Diese besteht aus drei Elementen. Das erste ist das Product-Service-Performance-Model, das ausgewählte DQ-Merkmale vier Quadranten zuteilt. Die Messung des DQ-Niveaus erfolgt dann mit Hilfe des zweiten Elements in Form einer Befragung der Endanwender nach deren Qualitätseinschätzungen. Als drittes Element von AIMQ werden mit Benchmark-Gap- und Role-Gap-Analyse eine anwendungsunabhängige sowie eine anwendungsabhängige Qualitätsanalyse der Messergebnisse vorgeschlagen. Problematisch bei diesem Vorgehen ist, dass die Messung der DQ auf einer subjektiven Qualitätseinschätzung anhand einer Befragung erfolgt. Dies ermöglicht keine fachlich interpretierbare und reproduzierbare Beurteilung des DQ-Niveaus (vgl. F6). Zudem ist eine zweckorientierte Messung der Qualität der Daten hinsichtlich deren konkreten Verwendung (vgl. F3) nicht vorgesehen. Stattdessen werden die subjektiven Einschätzungen mehrerer Nutzer vermischt, die i. d. R unterschiedliche Zwecke (mit den Daten) verfolgen.
Der Ansatz von Helfert [14] unterscheidet die Design- und die Konformitätsqualität [11]. Dabei bezeichnet die Designqualität den Grad der Übereinstimmung zwischen den Anforderungen der Datenanwender und der entsprechenden Repräsentation in der Spezifikation des Informationssystems. Die Konformitätsqualität drückt dagegen aus, in welchem Maße diese Spezifikation durch das Informationssystem eingehalten wird. Diese Unter-

[1] Folglich ist es nicht ausreichend, wenn sich die Metriken nur streng monoton wachsend bei verbesserter DQ im betrachteten Merkmal verhalten und ein ordinales Messsystem bilden wie bei [15].

scheidung ist im Hinblick auf eine Bewertung des DQ-Niveaus sinnvoll, da die subjektive Einschätzung der Übereinstimmung der Datenspezifikation mit dem Datenbedarf des Anwenders von der (objektivierbaren) Überprüfung der Konformität von vorhandenem und spezifiziertem Datenangebot getrennt wird. Den zentralen Aspekt bei Helfert stellt die Integration des DQ-Managements in die Metadatenverwaltung dar, die ein weitgehend automatisiertes und werkzeugunterstütztes Management der DQ ermöglichen soll. Die Qualitätsanforderungen sind dabei durch eine Regelmenge repräsentiert, die (automatisiert) überprüft wird, um Qualitätsaussagen abzuleiten. Insgesamt stellt Helfert jedoch keine konkreten Metriken dar, sondern hat den Anspruch das DQ-Management auf einer konzeptionellen Ebene zu beschreiben.

Neben diesen wissenschaftlichen Ansätzen sollen auch die bekannten praxisorientierten Konzepte von English und Redman kurz beschrieben werden. English verfolgt die Total Quality Data Management-Methode [7], die an die Konzepte des Total Quality Management angelehnt ist. Dabei führt er Vorgehensmuster zur Messung der Datendefinitions- und Architekturqualität (des Informationssystems) sowie der Qualität der Datenwerte und -repräsentation an. Obwohl das Verfahren in einer Reihe von Praxisprojekten Verwendung gefunden hat, findet sich hier jedoch leider kein allgemeines Vorgehen zur Messung der DQ. Redman verfolgt im Gegensatz zu English einen stark prozessorientierten Ansatz und kombiniert Messverfahren für gezielt ausgewählte Abschnitte im Informationsfluss mit dem Konzept der statistischen Qualitätskontrolle [23]. Einzelne Kennzahlen und Metriken werden dagegen nicht entwickelt.

Im Weiteren wird auf die Ansätze von Ballou et al. für das DQ-Merkmal Aktualität [2] sowie von Hinrichs [15] für die DQ-Merkmale Korrektheit und Aktualität eingegangen, da diese – im Gegensatz zur übrigen Literatur – konkrete Berechnungsvorschriften und Metriken zur Quantifizierung des DQ-Niveaus angeben. Die Vorgehensweise von Hinrichs ist aussichtsreich, da eine Bewertung des DQ-Niveaus mittels normierter (F1) und aggregierbarer (F4) Metriken erfolgt und Messverfahren zur Operationalisierung angeführt werden (F5). Allerdings wird bei genauer Betrachtung deutlich, dass mit den Metriken erhebliche Probleme einhergehen. Im Folgenden werden die vorgeschlagenen Metriken detaillierter vorgestellt.

Unter Korrektheit ist die Eigenschaft zu verstehen, in welchem Umfang die Attributwerte im Informationssystem den zugehörigen Ausprägungen der modellierten Realweltentität entsprechen – d. h., inwieweit die gespeicherten Datenwerte mit den realen Gegebenheiten übereinstimmen. Bei der Entwicklung der Metrik liegt folgende Definition zugrunde:

Sei w_I ein Attributwert im Informationssystem und w_R der entsprechende Wert des Attributs in der Realwelt. Sei zudem $d(w_I,w_R)$ ein domänenspezifisches auf das Intervall $[0;\infty]$ normiertes Abstandsmaß zur Bestimmung der Abweichung zwischen w_I und w_R. Als Beispiele für solche Abstandsmaße können das domänenunabhängige Abstandsmaß $d_1(w_I,w_R) := \begin{cases} 0 & \text{falls } w_I = w_R \\ \infty & sonst \end{cases}$, die Abstandsfunktion $d_2(w_I,w_R) := |w_I - w_R|$ für numerische, metrisch skalierte Attribute oder der Editierabstand und die Hamming-Distanz für Attribute des Typs String angeführt werden. Darauf basierend wird die Metrik für Korrektheit wie folgt definiert:

$$Q_{Korr.}(w_I,w_R) := \frac{1}{d(w_I,w_R)+1}$$

Dabei treten u. a. folgende konzeptionelle Probleme auf: Zum einen sind die resultierenden Werte nur sehr eingeschränkt fachlich interpretierbar (F6) – dies wird anhand des Beispiels in Abb. 2 deutlich, bei dem die Korrektheit des Attributs „Name“ für die Attributwerte „Meierhofre“ und „Mayerhofer“ bzw. „Mayr“ und „Wein“ untersucht wird. Als Abstandsmaß wird dabei auf die Hamming-Distanz zurückgegriffen, die als Bewertung die Anzahl der Positionen der beiden Strings liefert, an denen sich diese unterscheiden[2] – bei Verwendung alternativer Abstandsmaße, wie z. B. dem Editierabstand, treten die Probleme ebenso auf:

$$Q_{Korr.}(\text{"Meierhofre "},\text{"Mayerhofer "}) =$$

$$\frac{1}{d_{Ham.}(\text{"Meierhofre"},\text{"Mayerhofer"})+1} = \frac{1}{|\{2,3,9,10\}|+1} = \frac{1}{4+1} = 20{,}0\%$$

$$Q_{Korr.}(\text{"Mayr"},\text{"Wein"}) =$$

$$\frac{1}{d_{Ham.}(\text{"Mayr"},\text{"Wein"})+1} = \frac{1}{|\{1,2,3,4\}|+1} = \frac{1}{4+1} = 20{,}0\%$$

Abb. 2 Bestehende Metrik am Beispiel des Attributs „Name“

[2] Zugrunde liegende formale Definition der Hamming-Distanz für zwei Strings x und y mit gleicher Länge m ($|x|=|y|=m$): $d_{Ham.}(x,y)=|\{1\le i\le m \mid x[i]\neq y[i]\}|$. Hinweis: Bei Strings unterschiedlicher Länge kann der kürzere mit Dummy-Zeichen aufgefüllt werden, die jeweils als nicht übereinstimmend gelten.

Hier liefert die bestehende Metrik für beide Attributwerte das gleiche Resultat von 20,0%, obwohl der Name „Meierhofre“ bspw. im Rahmen einer Mailingkampagne noch (als „Mayerhofer“) identifiziert werden kann, wohingegen bei „Mayr“ und „Wein“ die erfolgreiche Zustellung einer Postsendung ausgeschlossen ist. Diese Schwäche beruht auf drei Ursachen: Erstens geht durch die Quotientenbildung im vorgeschlagenen Term die Interpretierbarkeit der Werte im Sinne von (F1) und (F2) verloren. Zweitens hängt die Größenordnung der Metrikergebnisse stark vom verwendeten Abstandsmaß und dem betrachteten Attribut ab. Darüber hinaus wird der Wertebereich von [0;1] i. d. R. nicht ausgeschöpft, da ein Metrikergebnis von 0 lediglich für den Fall resultiert, dass das Abstandsmaß den Wert ∞ liefert (bei Verwendung der Hamming-Distanz müssten somit z. B. unendlich viele Fehler vorliegen). Dadurch, dass weder eine absolute noch eine relative Veränderung der Metrikergebnisse interpretierbar ist und somit keine kardinale Skalierung vorliegt (F2), wird zudem die ökonomische Planung und Bewertung von DQ-Maßnahmen erschwert (bspw. hängt die Maßnahmenwirkung so mitunter von der Länge des qualitätsgesicherten Attributs ab). Tabelle 1 demonstriert letztgenannte Schwäche: Um den Wert für Korrektheit bspw. von 0,0 auf 0,5 zu verbessern, muss das entsprechende Abstandsmaß von ∞ auf 1,0 reduziert werden. Dagegen ist für eine Verbesserung von 0,5 auf 1,0 lediglich eine Reduzierung von 1,0 auf 0,0 nötig (d. h. bei der Verwendung der Hamming-Distanz müsste lediglich ein Fehler korrigiert werden). Offen bleibt, wie eine Verbesserung der Korrektheit um bspw. 0,5 zu interpretieren ist.

Tabelle 1 Veränderung der bestehenden Metrik und zugehörige Veränderung des Abstandsmaßes

Verbesserung der Korrektheit $Q_{Korr.}(w_I, w_R)$	**Notwendige Veränderung von** $d(w_I, w_R)$
$0{,}0 \rightarrow 0{,}5$	$\infty \rightarrow 1{,}0$
$0{,}5 \rightarrow 1{,}0$	$1{,}0 \rightarrow 0{,}0$

Neben Korrektheit wird das DQ-Merkmal Aktualität (für einen Literaturüberblick zu bestehenden Definitionen siehe [6]) betrachtet. Unter Aktualität ist die Eigenschaft der Gegenwartsbezogenheit der Daten zu verstehen, d. h. inwiefern die gespeicherten Werte den aktuellen Gegebenheiten in der Realwelt entsprechen und nicht veraltet sind. Im Gegensatz zur Messung der Korrektheit ist dabei ein Abgleich der Attributwerte mit den Gegebenheiten in der Realwelt nicht erforderlich. Folgende Metrik, die ähnliche Schwächen wie obige Metrik für Korrektheit aufweist, wird von Hinrichs vorgeschlagen (dort als Metrik für Zeitnähe bezeichnet [15]):

$$Q_{Akt.}(w, A) := \frac{1}{Update(A) \cdot Alter(w, A) + 1}$$

Alter(*w*,*A*) bezeichnet das Alter des Attributwerts *w* (als Differenz aus aktuellem Zeitpunkt und Befunddatum). *Update*(*A*) stellt die Updatehäufigkeit von Werten des Attributs *A* dar. Die Metrik verhält sich zwar in Bezug auf die berücksichtigten Parameter tendenziell richtig. Allerdings ist, wie bereits oben bei der Metrik für Korrektheit dargestellt, die Anforderung der kardinalen Skalierung (F2) aufgrund der Quotientenbildung verletzt. Zudem sind die Metrikergebnisse nicht fachlich interpretierbar (F6) – bspw. als Wahrscheinlichkeit dafür, dass der betrachtete Attributwert noch den aktuellen Gegebenheiten entspricht. So verschlechtert sich zwar das Metrikergebnis mit zunehmendem Alter des Attributwerts, jedoch haben die resultierenden Werte weder eine konkrete Aussagekraft (wie ist ein Wert von z. B. 0,5 zu deuten?) noch sind deren Veränderungen, bspw. aufgrund der Durchführung von DQ-Maßnahmen, interpretierbar (welche Konsequenzen hat eine Verbesserung der Metrik um z. B. 0,1?).
Die Metrik, die Ballou et al. definieren, stellt sich in einer angepassten, leicht vereinfachten Notation wie folgt dar [2]:

$$Q_{Akt.}(w, A) := \left\{ \max\left[1 - \frac{Alter(w, A)}{Gültigkeitsdauer(A)} \right], 0 \right\}^{s}$$

Alter(*w*,*A*) bezeichnet wiederum das Alter des Attributwerts *w*, das sich aus der Summe des Alters des Attributwerts zum Zeitpunkt der Datenerfassung im Informationssystem und der Differenz zwischen dem Zeitpunkt der DQ-Messung und dem Zeitpunkt der Datenerfassung ergibt. *Gültigkeitsdauer*(*A*) stellt einen Indikator für die Beständigkeit von Attributwerten des Attributs *A* dar. So ergibt sich bei relativ hoher *Gültigkeitsdauer*(*A*) ein relativ kleiner Quotient aus *Alter*(*w*,*A*) und *Gültigkeitsdauer*(*A*) und somit ein relativ hoher Wert der Metrik für Aktualität, vice versa. Über den Parameter *s*, den es von Fachexperten festzulegen gilt, kann beeinflusst werden, wie stark sich eine Veränderung des genannten Quotienten auf den Wert der Metrik auswirkt. So kann die Metrik auf das betrachtete Attribut und den speziellen Anwendungskontext angepasst werden.
Kritisch zu sehen ist, dass durch die Einführung des Exponenten *s* die fachliche Interpretierbarkeit der Metrikergebnisse (F6) sowie die kardinale Skalierung (F2) verloren gehen. Tabelle 2 demonstriert dies für *s*=2. Offen bleibt wiederum, wie eine Verbesserung der Metrik um bspw. 0,5 in Abhängigkeit vom Quotienten *Alter*(*w*,*A*)/*Gültigkeitsdauer*(*A*) zu interpretieren ist (eine Interpretationsmöglichkeit in Abhängigkeit von anderen Parametern ist ebenso nicht gegeben).

Tabelle 2 Veränderung der bestehenden Metrik und zugehörige Veränderung des Quotienten *Alter*(*w*,*A*)/*Gültigkeitsdauer*(*A*)

Verbesserung der Aktualität $Q_{Akt.}(w, A)$	**Notwendige Veränderung von** $\frac{Alter(w, A)}{Gültigkeitsdauer(A)}$
$0{,}0 \rightarrow 0{,}5$	$1 \rightarrow 0{,}29$
$0{,}5 \rightarrow 1{,}0$	$0{,}29 \rightarrow 0{,}0$

Insgesamt stellen Ballou et al. rein auf den funktionalen Zusammenhang ab, was zur Folge hat, dass eine fachliche Interpretierbarkeit der Metrik als Ganzes nicht gegeben ist. Eine Interpretation der Metrik als Wahrscheinlichkeit dafür, dass der betrachtete Attributwert noch den aktuellen Gegebenheiten entspricht, ist lediglich für $s=1$ auf Basis einer gängigen Verteilungsannahme möglich. So ist die Metrik in diesem Fall als Wahrscheinlichkeit bei zugrunde gelegter Gleichverteilung interpretierbar. Allerdings ist die Allgemeingültigkeit einer Metrik stark anzuzweifeln, der diese Annahme zugrunde liegt, da für die betrachtete Zufallsvariable eine feste maximale Lebensdauer sowie eine konstante (absolute) Verfallsrate unterstellt werden müssten. Konkret für den Kontext der DQ würde dies bedeuten: Für jedes Attribut existiert eine maximale Lebensdauer, die nicht überschritten werden kann. Dies ist z. B. für Attribute wie „Nachname“ oder „Geburtsdatum“ mehr als problematisch, da diese in der Realität weder eine fest begrenzte maximale Gültigkeitsdauer noch eine konstante Verfallsrate innerhalb dieses Betrachtungshorizontes aufweisen. Somit ist unmittelbar einsichtig, dass die Metrik höchstens für eine kleine Auswahl an Attributen Verwendung finden könnte.
Basierend auf den diskutierten Beiträgen, werden nachfolgend eigene Ansätze für die DQ-Merkmale Korrektheit und Aktualität entwickelt. Wie diese konkret angewendet werden können, wird insbesondere im übernächsten Kapitel am Beispiel eines Mobilfunkanbieters ausführlich demonstriert.

Entwicklung von neuen Datenqualitätsmetriken

Für das Merkmal Korrektheit sei w_I ein Attributwert im Informationssystem und w_R der entsprechende Attributwert in der Realwelt. Sei zudem $d(w_I,w_R)$ ein domänenspezifisches – im Gegensatz zu Hinrichs [15] auf das Intervall [0;1] normiertes – Abstandsmaß zur Bestimmung der Abweichung zwischen w_I und w_R. Als Beispiele können das domänenunabhängige Abstandsmaß $d_1(w_I, w_R) := \begin{cases} 0 & \text{falls } w_I = w_R \\ 1 & sonst \end{cases}$, die Abstandsfunktion $d_2(w_I, w_R) := \left(\frac{|w_I - w_R|}{\max\{|w_I|, |w_R|\}} \right)^{\alpha}$ mit $\alpha \in \mathrm{IR}^+$ für numerische, metrisch skalierte Attribute und auf das Intervall [0;1] normierte Varianten des Editierabstands oder der Hamming-Distanz für Attribute des Typs String angeführt werden. Die Metrik auf Attributwertebene definiert sich dann wie folgt:

$$Q_{Korr.}(w_I, w_R) := 1 - d(w_I, w_R) \tag{1}$$

Die Funktionsweise der Metrik wird am Beispiel der Attribute „Postleitzahl“ und „Hausnummer“ veranschaulicht, wobei die Abstandsfunktion $d_2(w_I,w_I)$ zugrunde gelegt wird: Zunächst gilt es, den Parameter α festzulegen, der die Sensibilisierbarkeit auf Attributwertebene gewährleistet (F3). Soll die Metrik stark auf kleine Abweichungen reagieren, so ist $\alpha<1$ zu wählen – bspw. bei der Untersuchung des Attributs „Postleitzahl“, da hier schon kleine Abweichungen dazu führen können, dass Kampagnenangebote nicht zugestellt werden können. Soll die Abstandsfunktion dagegen „toleranter“ gegenüber kleinen Abweichungen sein, ist $\alpha>1$ angebracht – wie z. B. beim Attribut „Hausnummer“, da eine Zustellung bei kleinen Abweichungen trotzdem möglich ist. Nach dieser Festlegung muss der Wert der Abstandsfunktion für w_I und w_R bestimmt und in Term (1) eingesetzt werden. Der resultierende Wert ist im Gegensatz zu bisheriger Metriken interpretierbar (F6) und der Wertebereich von [0;1] wird ausgeschöpft (F1). Dies wird anhand des bereits angeführten Beispiels in Abb. 3 illustriert. Als Abstandsmaß findet wiederum die Hamming-Distanz Verwendung, die mittels Division durch die Zeichenanzahl der längeren Zeichenkette auf das Intervall [0;1] normiert wird.

$$Q_{Korr.}(\text{"}Meierhofre\ \text{"},\text{"}Mayerhofer\ \text{"}) =$$

$$1 - d^{norm.}_{Ham.}(\text{"}Meierhofre\text{"},\text{"}Mayerhofer\text{"}) = 1 - \frac{|\{2,3,9,10\}|}{10} = 1 - \frac{4}{10} = 60{,}0\%$$

$$Q_{Korr.}(\text{"}Mayr\text{"},\text{"}Wein\text{"}) =$$

$$1 - d^{norm.}_{Ham.}(\text{"}Mayr\text{"},\text{"}Wein\text{"}) = 1 - \frac{|\{1,2,3,4\}|}{4} = 1 - \frac{4}{4} = 0{,}0\%$$

Abb. 3 Bewertung der Korrektheit des Attributs „Name“ anhand der entwickelten Metrik (Beispiel)

Dabei wird klar, dass die Metrik im Gegensatz zur bestehenden Metrik sehr wohl differenziert: Bei „Meierhofre“ und „Mayerhofer“ ergibt sich ein Metrikergebnis von 60,0%, wohingegen bei „Mayr“ und „Wein“ ein Wert von 0,0% (d. h. der Attributwert ist gar nicht korrekt und z. B. für die Durchführung einer Mailingkampagne unbrauchbar) resultiert, da die Zeichenketten keine Übereinstimmung aufweisen. Zudem ist die Metrik im Rahmen eines ökonomisch orientierten DQ-Managements einsetzbar, da auch die Forderung nach einer kardinalen Skalierung der Metrikergebnisse (F2) erfüllt ist (vgl. Tabelle 3). So muss, um das Metrikergebnis für Korrektheit bspw. um 0,5 zu verbessern, auch das entsprechende Abstandsmaß um 0,5 reduziert werden – unabhängig davon, ob die Metrik von 0,0 auf 0,5 oder von 0,5 auf 1,0 gesteigert werden soll.

Tabelle 3 Veränderung der entwickelten Metrik und zugehörige Veränderung des Abstandsmaßes

Verbesserung der Korrektheit $Q_{Korr.}(w_I, w_R)$	**Notwendige Veränderung von** $d(w_I, w_R)$
$0{,}0 \rightarrow 0{,}5$	$1{,}0 \rightarrow 0{,}5$
$0{,}5 \rightarrow 1{,}0$	$0{,}5 \rightarrow 0{,}0$

Die entwickelten Metriken sind im Falle eines relationalen Datenbankschemas einsetzbar und ermöglichen eine Bewertung auf Attributwert-, Tupel-, Relationen- sowie Datenbankebene. Der Anforderung der Aggregierbarkeit (F4) wird dadurch Rechnung getragen, dass die Metriken bottom up entwickelt werden – d. h. die Metrik auf Ebene n+1 (z. B. Korrektheit auf Tupelebene) basiert auf der Metrik auf Ebene n (Korrektheit auf Attributwertebene). Demzufolge wird die Qualitätsmetrik auf Tupelebene nun basierend auf der Metrik auf Attributwertebene definiert. Seien t ein Tupel mit Attributwerten $t.A_1,\ldots,t.A_{|A|}$ für die Attribute $A_1,\ldots,A_{|A|}$ und $e.A_1,\ldots,e.A_{|A|}$ die entsprechenden Ausprägungen der Realweltentität e. Des Weiteren sei die relative Bedeutung des Attributs A_i im Bezug auf Korrektheit jeweils mit $g_i \in [0;1]$ bewertet. Dann ergibt sich die Metrik auf Tupelebene zu:

$$Q_{Korr.}(t,e) := \frac{\sum_{i=1}^{|A|} Q_{Korr.}(t.A_i, e.A_i) g_i}{\sum_{i=1}^{|A|} g_i} \tag{2}$$

Auf Relationenebene kann die Korrektheit einer Relation oder eines Views R auf Basis des arithmetischen Mittels der Funktionswerte $Q_{Korr.}(t_j,e_j)$ der Metrik für die Tupel t_j aus R (j=1,…,$|T|$) definiert werden, sofern R eine nicht leere Relation und E die zugehörige Entitätenmenge in der Realwelt darstellen:

$$Q_{Korr.}(R,E) := \frac{\sum_{j=1}^{|T|} Q_{Korr.}(t_j, e_j)}{|T|} \tag{3}$$

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[3] – d. h. D lässt sich in paarweise überschneidungsfreie Relationen R_k zerlegen, so dass jedes Attribut in genau einem R_k enthalten ist (formal: $D=R_1\cup\ldots\cup R_{|R|}$ und $R_i\cap R_j=\varnothing\ \forall\ i\neq j$).

[3] Für den Fall, dass Schlüsselattribute in mehreren Relationen oder Views auftreten, sind diese ab dem zweiten Auftreten mit einer Gewichtung von null zu versehen, um eine mehrfache Berücksichtigung bei der Metrik für Korrektheit zu vermeiden. Die Anwendbarkeit der Metrik wird dadurch nicht eingeschränkt.

Weiter sei E der modellierte Ausschnitt der Realwelt, wobei zudem E_k die zu R_k zugehörige Entitätenmenge repräsentiert. Dann kann man die Korrektheit der Datenbank D auf Basis der Korrektheit der Relationen R_k (k=1,…,$|R|$) definieren:

$$Q_{Korr.}(D,E) := \frac{\sum_{k=1}^{|R|} Q_{Korr.}(R_k, E_k) g_k}{\sum_{k=1}^{|R|} g_k} \tag{4}$$

Über die Gewichtungsfaktoren $g_k \in [0;1]$ ist es im Gegensatz zu Hinrichs [15] – der auf ein ungewichtetes arithmetisches Mittel zurückgreift – möglich, die Bedeutung der Relationen für die Zielsetzung zu berücksichtigen (F3). Das ungewichtete arithmetische Mittel hat zur Folge, dass hinsichtlich des Ziels weniger relevante Relationen – im Vergleich zu wichtigen Relationen – gleich stark eingehen. Zudem ist in diesem Fall der Wert der Metrik von der konkreten Zerlegung der Datenbank in Relationen abhängig, was eine objektive Bewertung zusätzlich erschwert: Bspw. kommt der Relation R_k mit $k \neq 2$ bei der disjunkten Zerlegung $\{R_1,R_2,R_3,\ldots,R_n\}$ ein relatives Gewicht von $1/n$ zu, wohingegen dieselbe Relation bei Verwendung der disjunkten Zerlegung $\{R_1,R_2',R_2'',R_3,\ldots,R_n\}$ mit $R_2' \cup R_2'' = R_2$ und $R_2' \cap R_2'' = \emptyset$ nur mit dem Faktor $1/(n+1)$ eingeht.
Die Bewertung der Korrektheit ergibt sich somit direkt aus der dargestellten Metrik und den auf Attributwertebene definierten Abstandsmaßen. Dabei ist die Anwendung der Metrik für Korrektheit – im Gegensatz zur später vorgeschlagenen Metrik für Aktualität – i. d. R. nur für eine Stichprobe des gesamten Datenbestands praktikabel (z. B. könnte in Erwägung gezogen werden, für eine Stichprobe des Kundenstamms externe, aktuelle Adressdaten zuzukaufen und diese mit den eigenen Adressdaten abzugleichen). Dies liegt daran, dass hier im Sinne der Definition von Korrektheit ein Abgleich zwischen den Attributwerten im Informationssystem und der entsprechenden Ausprägung der Realweltentität unabdingbar ist. Für den gesamten Datenbestand ist dies jedoch nicht ohne Weiteres technisch, automatisiert und mit tolerierbarem Kostenaufwand möglich. Bei einer Anwendung der Metrik auf einen Teil der Daten können jedoch bei ausreichend großem Stichprobenumfang Rückschlüsse auf den gesamten Datenbestand gezogen und ein Schätzer für $Q_{Korr.}$ ermittelt werden (z. B. könnte man so über den genannten Zukauf von Adressdaten und den Abgleich mit den eigenen Kundendaten anhand der Metrik einen Schätzer für die Korrektheit des gesamten Adressdatenbestands ermitteln).
Neben der Korrektheit wird auch die Aktualität betrachtet. Unter Aktualität ist die Eigenschaft der Gegenwartsbezogenheit zu verstehen, d. h. inwiefern die im Informationssystem erfassten Werte den aktuellen Gegebenheiten in der Realwelt (noch) entsprechen. Die Bewertung basiert dabei auf wahrscheinlichkeitstheoretischen Betrachtungen, um die Interpretierbarkeit zu gewährleisten und eine automatisierte, reproduzierbare Analyse zu ermöglichen (F6). Aktualität kann hierbei als Wahrscheinlichkeit interpretiert werden, mit welcher bspw. der betrachtete Attributwert noch aktuell ist (F6). In der fachlichen Interpretierbarkeit liegt auch der Vorteil der Metrik im Vergleich zu existierenden Ansätzen. Sei A ein Attribut, w ein Attributwert und $Alter(w,A)$ das Alter des Attributwerts, das sich aus dem Zeitpunkt der Messung und dem Zeitpunkt der Datenerfassung errechnet. Sei zudem $Verfall(A)$ die (ggf. empirisch ermittelte) Verfallsrate von Werten des Attributs A – diese gibt an, wie viele Datenwerte des Attributs durchschnittlich innerhalb einer Zeiteinheit inaktuell werden. Dann ist die Metrik auf Attributwertebene wie folgt definiert:

$$Q_{Akt.}(w,A) := e^{-Verfall(A)\cdot Alter(w,A)} \tag{5}$$

Dabei stellt $Q_{Akt.}(w,A)$ unter der Annahme, dass die Gültigkeitsdauer der Datenwerte exponentialverteilt mit dem Parameter $Verfall(A)$ ist, die Wahrscheinlichkeit dar, mit welcher der Attributwert noch aktuell ist. $Verfall(A)$=0,2 ist hier bspw. so zu interpretieren, dass durchschnittlich 20% der Attributwerte des Attributs A innerhalb einer Zeiteinheit ihre Gültigkeit verlieren. Somit sind die Normierung der Metrik (F1) und die kardinale Skalierung der Metrikergebnisse (F2) gewährleistet. Bei der Exponentialverteilung handelt es sich um eine typische Lebensdauerverteilung, die sich insbesondere im Rahmen der Qualitätssicherung bewährt hat.
Bei Attributen wie z. B. „Geburtsdatum“ oder „Geburtsort“, die sich nie ändern, gilt dementsprechend $Verfall(A)$=0 und der Funktionswert der Metrik ergibt sich zu 1: $Q_{Akt.}(w,A) = e^{-Verfall(A)\cdot Alter(w,A)} = e^{-0\cdot Alter(w,A)} = e^{0} = 1$. Dies gilt auch für Attributwerte, die zum Betrachtungszeitpunkt neu erfasst werden (d. h. $Alter(w,A)$=0): $Q_{Akt.}(w,A) = e^{-Verfall(A)\cdot Alter(w,A)} = e^{-Verfall(A)\cdot 0} = e^{0} = 1$. Letzteres bedeutet, dass die Metrik auch bei erneuter Erfassung eines Attributwerts im Sinne einer Aktualisierung des bereits vorhandenen Attributwerts richtige Werte liefert.
Auf Tupel-, Relationen- und Datenbankebene kann die Metrik analog zu Korrektheit basierend auf der Metrik auf Attributwertebene definiert werden (F4). Im Folgenden wird neben der praktischen Anwendung veranschau-

licht, dass auch die zweckorientierte Sensibilisierbarkeit (F3) sowie die Operationalisierbarkeit der Metrik mittels Messverfahren (F5) gewährleistet sind.

Praktische Anwendung der Metrik für Aktualität

Die praktische Anwendung erfolgte im Rahmen des Kampagnenmanagements eines großen Mobilfunkanbieters. Aus Vertraulichkeitsgründen wurden die verwendeten Zahlen und Daten verändert und anonymisiert, wobei das Vorgehen sowie die Ergebnisse im Kern erhalten blieben. DQ-Probleme traten beim Unternehmen u. a. bei der Kundenansprache auf. Diese führten bspw. bei Mailingkampagnen dazu, dass oftmals keine korrekte und individuelle Kundenansprache möglich war, was sich in geringeren Erfolgsquoten niederschlug. Im Folgenden wird eine Kampagne zur Vermarktung einer Tarifoption betrachtet, d. h. Kunden mit dem Tarif „Mobil 500“ wird ein Angebot für einen Vertragswechsel zum Tarif „Mobil 1000“ unterbreitet, der wegen der längeren Vertragslaufzeit und der höheren Grundgebühr für den Mobilfunkanbieter profitabler ist.
Zunächst wurden die relevanten Attribute im Rahmen der Kampagne bestimmt. Diese waren „Name“, „Vorname“ und „Adresse“ (Straße, Hausnummer, Postleitzahl und Ort), um den Kunden das Angebot per Post zukommen zu lassen. Zudem war das Attribut “aktueller Tarif” essentiell, da das Angebot nur für Kunden mit dem Tarif „Mobil 500“ gültig sein sollte. Für diese Attribute musste dann in Workshops die relative Wichtigkeit im Hinblick auf die Zielsetzung bestimmt werden (F3). Als wichtigstes Attribut wurde „aktueller Tarif“ definiert, da nur Kunden mit dem Tarif „Mobil 500“ angesprochen werden sollten, d. h. das Attribut war Selektionskriterium. Es wurde daher mit einer relativen Wichtigkeit von 1,0 (Bezugsbasis für die anderen Attribute) versehen. Als zweitwichtigstes Attribut wurde „Adresse“ eingestuft, da eine Zustellung des Angebots ansonsten nicht möglich ist. Hier wurde jedoch nicht ebenfalls ein Wert von 1,0 vergeben, sondern nur 0,9, da Teile der Adresse – wie bspw. eine falsche Hausnummer – nicht zwingend notwendig für eine Zustellung des Angebots sind. Analog wurde dem Attribut „Name“ 0,9 zugeordnet, da bei einem Wechsel des Namens (bspw. bei Heirat) dem Zustellservice der „alte“ Name im Einzelfall noch bekannt ist. Demgegenüber wurde der Vorname des Kunden als unwichtiger eingestuft, da ein falscher Vorname ggf. zwar zu einer Verärgerung des Kunden führt, jedoch eine Kontaktierung nicht unmöglich macht. Um die Kundenbeziehungen nicht zu belasten, erhielt „Vorname“ dennoch den Wert 0,2. Danach musste aus dem aktuellen Betrachtungszeitpunkt und dem Zeitpunkt der letzten Datenerfassung, der beim Unternehmen als Metadatum hinterlegt ist, automatisiert das Alter der Attributwerte (d. h. $Alter(T.A_i,A_i)$) berechnet werden (F5). Zuletzt musste die Verfallsrate $Verfall(A_i)$ für die Attribute ermittelt werden. Im Anwendungsfall wurde wie folgt vorgegangen: Für die Attribute „Name“ und „Adresse“ wurden Daten des Statistischen Bundesamts zu Eheschließungen/Scheidungen bzw. zur Häufigkeit des Umzugs herangezogen. Dadurch konnten Verfallsraten von 0,02 für das Attribut „Name“ (d. h. pro Jahr wechseln durchschnittlich ca. 2% der Kunden ihren Namen) und 0,1 für das Attribut „Adresse“ geschätzt werden. Beim Attribut „Vorname“ wurde ein Wert von 0,0 angenommen, da sich der Vorname i. d. R. nicht ändert. Der Verfallsparameter für das Attribut „aktueller Tarif“ konnte aus eigenen Daten (Erfahrungswerte) zu 0,4 bestimmt werden – Tabelle 4 fasst die Werte zusammen:

Tabelle 4 Ermittlung der Aktualität anhand der entwickelten Metrik (Beispiel)

A_i	Name	Vorname	Adresse	Aktueller Tarif
g_i	0,9	0,2	0,9	1,0
$Alter(T.A_i,A_i)$ (in Jahren)	0,5	0,5	2,0	0,5
$Verfall(A_i)$ (in 1/Jahr)	0,02	0,00	0,10	0,40
$Q_{Akt.}(T.A_i, A_i) = e^{-Verfall(A_i)\cdot Alter(T.A_i,A_i)}$	0,99	1,00	0,82	0,82

Im Beispiel ergibt sich der Wert der Metrik auf Tupelebene durch Aggregation der Ergebnisse auf Attributwertebene zu:

$$Q_{Akt.}(T, A_1,..., A_4) = \frac{0{,}99 \cdot 0{,}9 + 1{,}00 \cdot 0{,}2 + 0{,}82 \cdot 0{,}9 + 0{,}82 \cdot 1}{0{,}9 + 0{,}2 + 0{,}9 + 1{,}0} \approx 0{,}882$$

So liefert die Metrik einen Wert von 88,2%, d. h. das betrachtete Tupel ist für den konkreten Anwendungsfall *Vermarktung einer Tarifoption* zu 88,2% aktuell. Bevor derartige Berechnungen für die aktuelle Kampagne erfolgten, wurde zunächst eine ca. 3 Monate zurückliegende, ähnliche Kampagne analysiert, in der insgesamt 82.000 Kunden angeschrieben wurden, um ihnen einen Tarifwechsel anzubieten. Damals ergab sich eine durchschnittliche Erfolgsquote von ca. 8,5%, d. h. ca. 7.000 Kunden konnten vom Tarifwechsel überzeugt werden. Für alle Kunden dieser Kampagne wurde das Metrikergebnis für Aktualität berechnet. Danach wurden die Kunden in Abhängigkeit vom Ergebnis zu Gruppen zusammengefasst, d. h. jeder Kunde wurde je nach Wert den Intervallen [0,0;0,1],]0,1;0,2], …,]0,9;1,0] zugeordnet. Für jedes Intervall/Gruppe wurde dann im nächsten Schritt ermit-

telt, wie viele Kunden das Angebot (in der zurückliegenden Kampagne) angenommen haben (Erfolgsquote in Abhängigkeit vom Metrikergebnis für Aktualität). Die Auswertung zeigt Abb. 4:

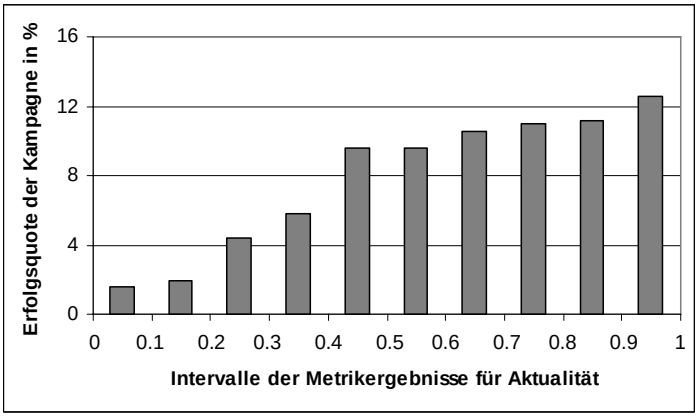

Abb. 4 Erfolgsquoten der zurückliegenden Kampagne in Abhängigkeit vom Metrikergebnis für Aktualität

Die Berechnungen zeigen, dass die Erfolgsquote umso höher ist, je höher die Aktualität der Kundenattribute ist. Sie beträgt im Intervall]0,2;0,3] nur 4,4%, wohingegen sie im Intervall]0,9;1,0] immerhin bei 12,6% liegt. Dies verwundert nicht, da bspw. Kunden mit einer veralteten, falschen Adresse gar nicht die Möglichkeit haben, das Angebot anzunehmen. Interessant werden diese Auswertungen nunmehr in Zusammenhang mit der neuen Kampagne, bei der insgesamt 143.000 Kunden angeschrieben werden könnten (alle Kunden mit dem Tarif „Mobil 500“). Legt man die Aufteilung der Erfolgsquoten in Abhängigkeit vom Metrikergebnis für Aktualität aus Abb. 4 zugrunde, so ist es unter Kosten-Nutzen-Gesichtspunkten nicht sinnvoll, Kunden mit einem Metrikergebnis für Aktualität kleiner als 0,3 überhaupt anzuschreiben. Bspw. befinden sich im Intervall]0,2;0,3] 12.500 Kunden, für die jedoch bei einer (erwarteten) Erfolgsquote von nur 4,4% (somit ca. 550 erwartete erfolgreiche Kontakte) die Kosten für die Unterbreitung des Angebots größer sind als der resultierende Mehrerlös durch den Tarifwechsel. Erst ab dem Intervall]0,4;0,5] übersteigen die (erwarteten) Mehrerlöse die Mailingkosten (siehe Abb. 5 rechts). Berücksichtigt man demnach nur Kunden mit einem Ergebnis größer als 0,4, so lässt sich die Effizienz der Kampagne steigern, da sowohl die Kosten reduziert als auch die (erwartete) Erfolgsquote verbessert werden.

Ein ökonomisch orientiertes DQ-Management will jedoch noch mehr. Zwar konnte durch die Auswertung die Effizienz erhöht werden, allerdings ist es unbefriedigend, dass Kunden – die ggf. den Tarif wechseln würden – das Angebot mangels Zustellung (bspw. veraltete Adresse) nicht akzeptieren können. Deswegen wird untersucht, inwieweit der Zukauf externer Daten als DQ-Maßnahme helfen könnte (vgl. DQ-Regelkreis). Unternehmen wie die Deutsche Post bieten bspw. aktuelle Adressdaten an[4]. So können die eigenen, vorliegenden Adressdatenbestände (zumindest teilweise) mit den aktuellen Daten der Deutschen Post abgeglichen und bei Bedarf aktualisiert bzw. ersetzt werden. Dadurch kann sichergestellt werden, dass die eigenen Adressdatenbestände, die für die Unterbreitung des Kampagnenangebotes genutzt werden, aktuell sind und der Kunde auf dem Postweg tatsächlich erreicht wird. Somit stellt sich die Frage, wie diese Maßnahme eingesetzt werden soll, da der Zukauf von Adressdaten einerseits natürlich Kosten verursacht, andererseits aber auch die Erfolgsquote von Kampagnen verbessert (da das Kampagnenangebot verhältnismäßig mehr Kunden tatsächlich erreicht). Beim Mobilfunkanbieter wird folgendermaßen vorgegangen: Zuerst wird pro Intervall berechnet, welche Kosten für den Zukauf der Adressen der Kunden (des jeweiligen Intervalls) anfallen würden. Dies ist unproblematisch, da die Anzahl der Kunden pro Intervall bekannt ist und Unternehmen wie die Deutsche Post einen Fixpreis je aktualisierter Adresse berechnen. Dieser Kostenkalkulation sind die resultierenden Erlöse infolge einer höheren Erfolgsquote (Angebot kann den Kunden nun zugestellt werden) gegenüber zu stellen. Hierzu ist im ersten Schritt das Ausmaß der Steigerung der DQ infolge des Adresszukaufs (für jedes betrachtete Intervall) zu berechnen, was mit Hilfe der obigen Formel $Q_{Akt.}(T.A_i,A_i)$ möglich ist. Anhand der neu ermittelten Metrikergebnisse lässt sich dann die verbesserte Erfolgsquote der Kampagne auf Basis der Erfolgsquoten der vergangenen Kampagne (für jedes Intervall) abschätzen (vgl. Abb. 5 links). Mit dieser gehen wiederum zusätzliche Erlöse einher, die es mit den Kosten des Adresszukaufs zu vergleichen gilt. Das rechte Diagramm aus Abb. 5 illustriert diesen Vergleich, wobei die relevanten Berechnungen zunächst noch detaillierter am Beispiel des Intervalls]0,2;0,3] erläutert werden. So sind ohne Adresskauf bei den 12.500 Kunden dieses Intervalls bei einer geschätzten Erfolgsquote von 4,4% insgesamt 550 erfolgreiche Kundenkontakte zu erwarten. Folglich stehen den Erlösen der Kampagne von 11.000 Euro (bei einem Erlös von 20 Euro pro Tarifwechsel) Mailingkosten von 15.625 Euro (bei Kosten von 1,25 Euro pro Mailing) gegenüber, was (ohne Adresskauf) einen negativen Wert für „Erlöse abzüglich Kosten“ in Höhe von

[4] vgl. z. B. Internetauftritt der Deutschen Post Direkt GmbH

-4.625 Euro zur Folge hat. Demgegenüber können im Falle des Adresskaufs bei einer geschätzten verbesserten Erfolgsquote von 9% zusätzlich 575 Kunden erfolgreich kontaktiert werden. Dabei sind den so zusätzlich resultierenden Erlösen von 11.500 Euro Kosten für den Adresskauf in Höhe von 5.000 Euro (0,4 Euro pro Kunde) gegenüber zu stellen. Insgesamt hat der Adresskauf somit zusätzliche „Erlöse abzüglich Kosten“ in Höhe von 6.500 Euro zur Folge.

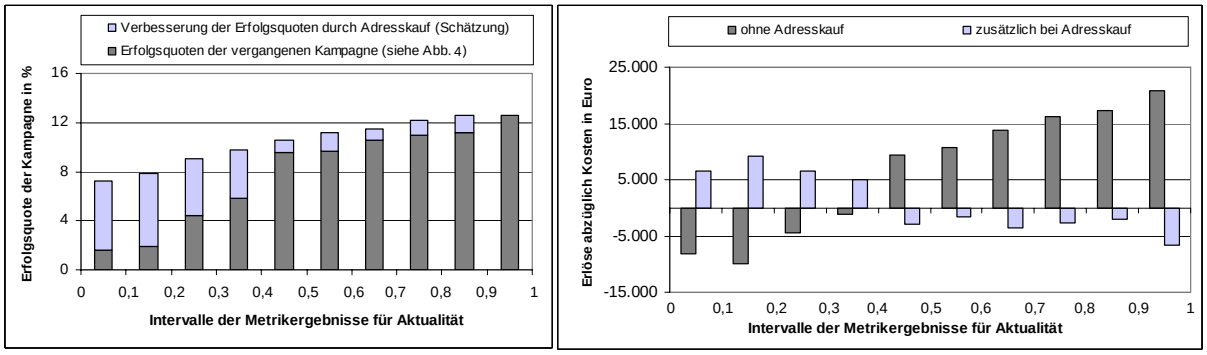

Abb. 5 Erfolgsquoten und Erlöse abzüglich Kosten der Kampagne in Abhängigkeit vom Metrikergebnis für Aktualität

Abb. 5 zeigt dabei zweierlei: Zum einen ist nur in den Intervallen des Bereichs]0,4;1,0] die Durchführung der Kampagne ohne Adresskauf ökonomisch sinnvoll. Dieser ist zudem auch nicht ratsam, da die Kosten des Zukaufs größer sind als die (erwarteten) zusätzlichen Erlöse (negativer Wert für Erlöse abzüglich Kosten). Dagegen ist bei den Kunden, die in die Intervalle des Bereichs [0,0;0,4] fallen, – ohne den Zukauf von Adressen zu berücksichtigen – die Kampagnendurchführung nicht sinnvoll. Jedoch zeigt sich, dass hier ein Adresszukauf zu positiven Zuwächsen führt (zusätzliche Erlöse abzüglich Kosten des Adresszukaufs sind größer als 0). Allerdings sind nur in den Intervallen]0,2;0,3] und]0,3;0,4] die gesamten (erwarteten) Erlöse größer als die Summe der Mailingkosten und der Kosten des Adresszukaufs. Insofern ist nur für die ca. 25.400 Kunden dieser Intervalle ein Adresszukauf rentabel, falls nur die in naher Zukunft generierbaren Erlöse in den Business Case einbezogen werden, was einer sehr vorsichtigen Kalkulation entspricht. Andererseits ist gerade die Erlöswirkung von DQ-Maßnahmen in (viel) späteren Perioden oftmals schwer gegenüber den Investitionsentscheidern zu argumentieren.
Neben dem kurzen Beispiel für die Anwendung der Metrik, bei dem die Kosten für die Kampagne und die Durchführung der DQ-Maßnahmen gesenkt werden konnten, wurde eine Reihe weiterer Analysen durchgeführt, um Kosten zu sparen oder Erlöse zu generieren. Insgesamt konnte der Mobilfunkanbieter durch die Anwendung der Metriken einen direkten Zusammenhang zwischen den Ergebnissen der DQ-Messung und den Erfolgsquoten von Kampagnen herstellen. So konnte der Prozess der Kundenselektion für Kampagnen deutlich verbessert werden. Zudem konnten auf Basis der Metriken DQ-Maßnahmen gezielter eingesetzt sowie damit einhergehende ökonomische Folgen besser abgeschätzt werden.

Zusammenfassung

Im Beitrag wurde die Fragestellung aufgegriffen, wie DQ adäquat quantifiziert werden kann. Ziel war, neue Metriken für die DQ-Merkmale Korrektheit und Aktualität vorzustellen, die eine objektive, zielgerichtete und weitgehend automatisierbare Messung auf unterschiedlichen Aggregationsebenen (Attributwerte, Tupel, etc.) ermöglichen. In Zusammenarbeit mit einem Mobilfunkanbieter konnten die Metriken angewendet und auf ihre Eignung für den Einsatz in der Praxis untersucht werden. Dabei wurde im Gegensatz zu bestehenden Ansätzen z. B. insbesondere Wert auf eine kardinale Skalierung gelegt, um auch ökonomische DQ-Betrachtungen zu unterstützen. Die Metriken ermöglichen somit eine Quantifizierung der DQ und bilden die Basis für eine Reihe ökonomischer Analysen. So können zukünftige Auswirkungen auf die DQ, wie z. B. zeitlicher Verfall oder die Durchführung von DQ-Maßnahmen, untersucht und damit ex ante Planungswerte mit ex post Messwerten verglichen werden.
Demgegenüber ist die Annahme einer exponentialverteilten Gültigkeitsdauer der Attributwerte bei der Entwicklung der Metrik für Aktualität durchaus kritisch zu sehen. Ob diese Annahme für die konkrete Anwendung gerechtfertigt werden kann, bleibt somit für den Einzelfall zu untersuchen. Falls dem nicht so sei, kann die Metrik analog zu oben ebenfalls basierend auf anderen Wahrscheinlichkeitsverteilungen definiert werden (die gewünschten Anforderungen an die Metrik bleiben dabei erfüllt). Darüber hinaus ist eine Ausweitung der Metriken auf weitere DQ-Merkmale notwendig. Dies stellt ebenso weiteren Forschungsbedarf dar, wie Ansätze zur Aggregation der Bewertungen für verschiedene DQ-Merkmale zu einem Gesamtqualitätswert (vgl. [20]). Parallel wird weiter an modellbasierten Ansätzen zur ökonomischen Planung von DQ-Maßnahmen gearbeitet (vgl. z. B.

[13]), für deren Operationalisierung Metriken für DQ unbedingt erforderlich sind. Die im Beitrag vorgestellten Ansätze bilden hierfür eine geeignete Grundlage.

Literatur

1. Alt, G.: Sehr geehrter Frau Müller – Falsche Daten sind nicht nur peinlich, sondern verursachen auch hohe Kosten. FAZ 244, B2 (2003)
2. Ballou, D. P., Wang, R. Y., Pazer, H., Tayi, G. K.: Modeling information manufacturing systems to determine information product quality. Management Science 44 (4), 462-484 (1998)
3. Bamberg, G., Baur, F., Krapp, M.: Statistik. München, Wien: Oldenburg 2007
4. Batini, C., Scannapieco, M.: Data Quality: Concepts, Methods and Techniques. Berlin: Springer 2006
5. Campanella, J.: Principles of quality cost. Milwaukee: ASQ Quality Press 1999
6. Cappiello, C., Francalanci, C., Pernici, B.: Time-Related Factors of Data Quality in Multichannel Information Systems. Journal of Management Information Systems 20 (3), 71-91 (2004)
7. English, L.: Improving Data Warehouse and Business Information Quality. New York: Wiley 1999
8. Eppler, M. J.: Managing Information Quality. Berlin: Springer 2003
9. Even, A., Shankaranarayanan, G.: Value-Driven Data Quality Assessment. Proceedings of the 10th International Conference on Information Quality. Cambridge: 2005
10. Feigenbaum, A. V.: Total quality control. New York: McGraw-Hill Professional 1991
11. Heinrich, B.; Helfert, H.: Analyzing Data Quality Investments in CRM – a model based approach. Proceedings of the 8th International Conference on Information Quality. Cambridge: 2003
12. Heinrich, B., Kaiser, M., Klier, M.: Metrics for measuring data quality - Foundations for an economic oriented management of data quality. Proceedings of the 2nd International Conference on Software and Data Technologies (ICSOFT). Barcelona: 2007
13. Heinrich, B., Klier, M.: Ein Optimierungsansatz für ein fortlaufendes Datenqualitätsmanagement und seine praktische Anwendung bei Kundenkampagnen. Zeitschrift für Betriebswirtschaft 76 (6), 559-587 (2006)
14. Helfert, M.: Proaktives Datenqualitätsmanagement in Data-Warehouse-Systemen - Qualitätsplanung und Qualitätslenkung. Berlin: Buchholtz, Volkhard, u. Thorsten Pöschel 2002
15. Hinrichs, H.: Datenqualitätsmanagement in Data Warehouse-Systemen, Dissertation der Universität Oldenburg. Oldenburg: 2002
16. Jarke, M., Vassiliou, Y.: Foundations of Data Warehouse Quality – A Review of the DWQ Project. Proceedings of the 2nd International Conference on Information Quality. Cambridge 1997
17. Lee, Y. W., Strong, D. M., Kahn, B. K., Wang, R. Y.: AIMQ: a methodology for information quality assessment. Information & Management 40, 133-146 (2002)
18. Machowski, F., Dale, B. G.: Quality costing: An examination of knowledge, attitudes, and perceptions. Quality Management Journal 3 (5), 84-95 (1998)
19. Matzer, M.: Datenqualität frisst die Hälfte des Data-Warehouse-Etats. Computerzeitung 3, 12 (2004)
20. Naumann F.: Aktuelles Schlagwort: Datenqualität. Informatik Spektrum 30 (1), 27-31 (2007)
21. Naumann F., Rolker, C.: Assessment Methods for Information Quality Criteria. Proceedings of the 5th International Conference on Information Quality. Cambridge: 2000
22. Pipino, L., Lee, Y., Wang, R.: Data quality assessment. Communications of the ACM 45 (4), 211–218 (2002)
23. Redman, T. C.: Data Quality for the Information Age. Norwood: Arctech House 1996
24. Redman, T. C.: The Impact of Poor Data Quality on the Typical Enterprise. Communications of the ACM 41 (2), 79-82 (1998)
25. Shank, J. M., Govindarajan, V.: Measuring the cost of quality: A strategic cost management perspective. Journal of Cost Management 2 (8), 5-17 (1994)
26. Strong, D. M.Lee, Y. W., Wang R. Y.: Data quality in context. Communications of the ACM 40 (5), 103-110 (1997)
27. Wang, R. Y., Storey, V. C., Firth, C. P.: A Framework for analysis of data quality research. IEEE Transaction on Knowledge and Data Engineering 7 (4), 623-640 (1995)
28. Wang, R. Y., Strong, D. M.: Beyond accuracy: What data quality means to data consumers. Journal of Management Information Systems 12 (4), 5–34 (1996)