Datenqualität

Seminar Informationsintegration and Informationsqualität

Ausarbeitung von Siegfried Wirth
Betreuer: Prof. Dr.-Ing. Dr. h. c. Theo Härder

Technische Universität Kaiserslautern
Lehrgebiet Datenverwaltungssysteme
Sommersemester 2006

Zusammenfassung Integrierte heterogene Informationssysteme werden in verschiedenen Anwendungsbereichen immer wichtiger. Bei der Integration von Daten aus verschiedenen Datenquellen ist es notwendig, die Qualität der Daten zu kennen oder beurteilen zu können. Nur dann ist es möglich, aus den zur Verfügung stehenden autonomen Datenquellen ein Ergebnis zusammen zu stellen, das für den Nutzer seinen Zweck erfüllt, nämlich belastbare Daten und Informationen zu liefern. In dieser Ausarbeitung werden ausgehend von einem intuitiven Qualitätsbegriff relevante Qualitätskriterien für Daten und ihre Quellen sowie Probleme bei ihrer Erhebung identifiziert. Darauf aufbauend werden Möglichkeiten vorgestellt, wie einzelne Kriterien bewertet und gewichtet werden können. Darüber hinaus werden Verfahren vorgestellt, wie Qualitätskriterien bei der Anfragebearbeitung berücksichtigt werden, um Datenquellen auszuwählen. Bei der Betrachtung dieser theoretischen Fragestellungen wird, an Hand von Beispielen, auf praktische Anwendungen verwiesen. Ausgangspunkt für alle Betrachtungen ist eine Mediator-Wrapper-Architektur, die skizziert wird.

Inhaltsverzeichnis

1 Einleitung

In diesem Kapitel wird die Bedeutung der Datenqualität in integrierten heterogenen Informationssystemen motiviert. Im Rahmen dieser Ausarbeitung nennen wir integrierte heterogene Informationssysteme nur noch heterogene Informationssysteme, da es für dieses Thema klar ist, dass das Ziel eines jeden solchen Systems die Integration der heterogenen Datenquellen ist. Dazu betrachten wir auf einer allgemeinen Ebene zunächst die besondere Bedeutung der Datenqualität. Abschließend folgen eine inhaltliche Abgrenzung dieser Ausarbeitung und ein Ausblick auf die weiteren Kapitel.

1.1 Bedeutung der Datenqualität in heterogenen Informationssystemen

In heterogenen Informationssystemen stehen meistens sehr viele Datenquellen zur Verfügung. Als Beispiel können wir jede vorhandene Webseite im Internet als eine Datenquelle betrachten. In anderen Anwendungen ist eine so große Anzahl von Datenquellen nicht direkt ersichtlich, aber für die meisten Anwendungsdomänen gilt doch, dass es eine mehr als ausreichende Anzahl gibt. Damit stellt sich für ein heterogenes Informationssystem die Frage, welche Datenquellen für eine konkrete Anfrage ausgewählt werden sollen. Alle abzufragen wird oftmals zu lange dauern. Außerdem wird ein gutes Ergebnis weniger von einer großen Quantität erreicht, sondern vielmehr durch Auswahl der qualitativ hochwertigsten Daten für die aktuelle Anfrage. In diesem Kontext ist die Bearbeitung des Themas Datenqualität in dieser Ausarbeitung zu verstehen. Wir wollen aufzeigen, wie es möglich ist, eine solche qualitativ hochwertige Auswahl zu treffen. Nur wenn es dem Informationssystem gelingt, eine solche Auswahl zu treffen, wird der Nutzer mit verlässlichen Daten in angemessener Zeit versorgt.

1.2 Inhaltliche Abgrenzung und Ausblick

Aus der Bedeutung der Datenqualität in heterogenen Informationssystemen ergibt sich als Motivation die Fragestellung, wie eine gute Auswahl von Datenquellen vorgenommen werden kann. Dazu ist es notwendig den Begriff der Datenqualität zu untersuchen und zu systematisieren sowie Methoden zur Erhebung und Verarbeitung von Qualitätsdaten vorstellen. Diese beiden Fragestellungen bilden den Kern dieser Ausarbeitung.

Wir beschäftigen uns in Kapitel 2 zunächst mit Grundlagen für diesen Themenbereich. Wir befassen uns etwas allgemeiner mit dem Begriff Qualität, um unsere Sichtweise des Themenbereichs klarer abzugrenzen und ein intuitives Verständnis für Datenqualität zu entwickeln. Darüber hinaus stellt Kapitel 2 wichtige Anwendungsdomänen vor, die wir teilweise im Verlauf der Ausarbeitung als Beispiele verwenden. Dabei berühren wir einige erweitere Aspekte, denen wir in dieser Ausarbeitung nicht weiter nachgehen können, die wir aber für wichtig halten, um das Thema in einen größeren Rahmen einzuordnen und um auf unterschiedliche Perspektiven hinzuweisen. Außerdem legen wir uns auf ein

Architekturmodell für heterogene Informationssysteme fest, das wir zu diesem Zweck skizzieren.

In Kapitel 3 stellen wir Qualitätskriterien und damit eine mögliche Systematisierung vor. Daraus ergibt sich die Fragestellung für Kapitel 4, das sich dem Problem zuwendet, diese Qualitätsdaten zu erheben. Dazu identifizieren wir die wesentlichen Quellen für Qualitätsdaten, was eine weitere Systematisierung darstellt. Nach diesen beiden Kapiteln nutzen wir die daraus gewonnenen Informationen in Kapitel 5 unter Zuhilfenahme mathematisch geprägter Methoden, um ein Qualitätsurteil für eine Datenquelle zu finden oder es zu ermöglichen, Datenquellen bezüglich ihrer Datenqualität zu vergleichen und wenden uns damit der Verarbeitung von Qualitätsdaten zu.

Abschließend kommen wir auf die ursprüngliche Fragestellung zurück und zeigen in Kapitel 6 auf, wie aus solchen Qualitätsurteilen Entscheidungen getroffen werden können, um Datenquellen für die Beantwortung einer konkreten Anfrage auszuwählen.

Dabei beschränken wir uns auf die Aspekte, die direkt mit der Datenqualität im Zusammenhang stehen. Fragen, die sich mit der konkreten Zusammenführung verschiedener Datensätze, der Erkennung von Duplikaten oder der Bereinigung fehlerhafter Daten beschäftigen, sind nicht Inhalt dieser Ausarbeitung.

Kapitel 7 fasst die wesentlichen Erkenntnisse zusammen und bietet einen abschließenden Überblick über die behandelten Themen.

2 Grundlagen und Architekturmodell

In diesem Kapitel befassen wir uns mit Grundlagen im Bereich der Datenqualität. Dazu gehören eine Präzisierung des Begriffes der Datenqualität ausgehend von dem allgemeinen Begriff Qualität und wichtige Anwendungsdomänen heterogener Informationssysteme. Diese Anwendungen bieten uns die Möglichkeit, einige weitergehende Perspektiven und allgemeine Beobachtungen anzudeuten. Dieses Kapitel schließen wir mit der Vorstellung eines Architekturmodells für heterogene Informationssysteme ab, das wir allen weiteren Betrachtungen im Rahmen dieser Ausarbeitung zu Grunde legen.

2.1 Qualität

Der Begriff der Datenqualität setzt voraus, dass wir zunächst kurz unsere Sichtweise des Qualitätsbegriffes darlegen und aufzeigen, wie dieser Qualitätsbegriff auf Datenqualität anzuwenden ist. Qualität verstehen wir nicht aus der Sicht des Anbieters eines Produktes oder einer Datenquelle, sondern aus der Sicht des Nutzers[1] des Produktes oder der Datenquelle. Dies schließt Nutzer innerhalb der Organisation, in der die Datenquelle existiert, ebenso ein wie solche, die von außen über Netzwerke, insbesondere das Internet, diese Datenquelle nutzen. Damit

[1] Auch wenn hier der Benutzer noch nicht personalisiert ist, sondern auch ein technisches System sein kann, wollen wir an dieser Stelle darauf hinweisen, dass wir in dieser Ausarbeitung auf geschlechtsspezifische Bezeichnungen verzichten.

hat der Qualitätsbegriff bereits zwei Ebenen: Zum einen Qualitätsmerkmale der Datenquelle an sich, wie z.B. die Antwortzeit; die wenigsten Benutzer werden bereit sein, mehrere Tage auf die Antwort einer Datenquelle zu warten. In dieser ersten groben Annäherung zählen wir hierunter auch sehr subjektive Werte einer Datenquelle, z.B. in welchem Maß der Nutzer dem Anbieter der Datenquelle Vertrauen entgegen bringt. Zum anderen gibt es Qualitätsmerkmale, die sich auf die gelieferten Daten beziehen, wie z.B., ob die Daten korrekt im Sinne des Nutzers sind; so werden einige Benutzer mit gewissen Unschärfen zufrieden sein, andere Benutzer wollen nur sichere und überprüfte Daten. Zu diesen Qualitätsmerkmalen wollen wir an dieser Stelle die Präsentation der Daten rechnen.

Für nicht verteilte, homogene Datenquellen, insbesondere für Anwendungen wie OLTP-Anwendungen (Online Transaction Processing), die auf Datenbanken aufbauen, waren bisher die Antwortzeit und der Durchsatz die wesentlichen Qualitätskriterien. Dieser Qualitätsbegriff ist leicht zu systematisieren und eine Datenerhebung für die Qualitätskriterien kann automatisch, vor allem im Rahmen von standardisierten Benchmarks, vorgenommen werden. Die vorherigen Betrachtungen haben gezeigt, dass diese Qualitätskriterien für heterogene Datenquellen weiterhin eine Rolle spielen, aber ergänzt werden müssen.

2.2 Qualität und Nutzeranforderungen

Die Anforderungen des Nutzers bestimmen die Dimensionen der Qualität. Daraus folgt nach Zink [Zi94] "zwingend, dass es ohne Spezifizierung der Anforderungen keine Qualität geben kann." Diesen Punkt werden wir ausführlich betrachten, da es zunächst nicht klar ist, wie man die Qualität von Daten oder von Datenquellen systematisieren, beschreiben und somit in diesem Sinne spezifizieren kann. Je mehr kostenpflichtige Datenanbieter verfügbar sind, desto wichtiger wird die daraus von Zink [Zi94] entwickelte erweiterte Qualitätsdefinition: "Qualität ist die Erfüllung von (vereinbarten) Anforderungen zur dauerhaften Kundenzufriedenheit." Dieser sehr weitgehende Blickwinkel ist für diese Arbeit in soweit zu relativieren, dass viele Informationsdienste kostenlos im Internet verfügbar sind und es noch keinen ausgeprägten Markt für verteilte Datenquellen gibt. An dieser Stelle wird die von Masing in [Ma99] postulierte Annahme relativiert, dass "die Nichterfüllung einer Qualitätsanforderung [...] das Erzeugnis fehlerhaft macht", in dem Sinne, dass der Nutzer, der in dieser Situation vor der Frage steht, ob er schlechte oder keine Daten haben möchte, in gewissen Grenzen zu Kompromissen bereit sein wird. Daraus wird deutlich, dass die Gewichtung der einzelnen Dimensionen durch den Nutzer von erheblicher Bedeutung ist, d.h., auf welche Dimensionen von Qualität er nicht zu verzichten bereit ist, welche ihm besonders wichtig sind und welche ihm nicht so wichtig sind. Damit kann eine Übererfüllung einer wichtigen Anforderung unter Umständen einen Mangel einer weniger wichtigen ersetzen.

Für zukünftige kommerzielle Anwendungen sollte aber ein ähnlich umfassender Qualitätsbegriff wie von Zink zu Grunde gelegt und trotzdem die angedeutete Gewichtung berücksichtigt werden; je mehr Anbieter es gibt, desto weniger fallen die Relativierungen und Einschränkungen ins Gewicht.

Wir fassen zusammen, dass wir Qualität stets auf den Nutzer und seine Erwartungen beziehen und sowohl die konkreten Daten als auch die Datenquelle betrachten werden.

2.3 Anwendungsdomänen heterogener Informationssysteme

Verschiedene Anwendungsbereiche nutzen bereits in großem Umfang heterogene Informationssysteme. Die bekanntesten und vermutlich ältesten Anwendungen von solchen Systemen sind Metasuchmaschinen im Internet, die verschiedene Suchmaschinen abfragen und deren Einzelergebnisse zu einem Gesamtergebnis koordinieren.

In der Molekularbiologie spielen heterogene Informationssysteme eine immer größere Rolle im Bereich der Genomforschung. Viele Forschungseinrichtungen erforschen von verschiedenen Lebewesen Gene, Regionen von Genen, damit zusammenhängende Merkmalsausprägungen, insbesondere Krankheiten, und stellen diese Informationen über Webdienste zur Verfügung. Eine bekannte Datenbank in diesem Bereich ist die des Human Genome Project, in dem das menschliche Genom nahezu vollständig erfasst wurde.

Eine weitere, mehr kommerzielle Anwendung heterogener Informationssysteme sind Börsenkurse. Viele Banken und Börsen selbst bieten Kursinformationen über das Internet an. Die Informationssysteme von Börsen bieten dabei meistens nur Informationen über diejenigen Aktien, die an dieser Börse gehandelt werden. Eine Sammlung dieser Informationen aus allen Quellen zu einem aktuellen Stand der Kurse aller Aktien ist eine Herausforderung, zumal die teilweise angebotenen Zusatzinformationen, wie Firmeninformationen, Kursentwicklung, Pressemeldungen etc. ebenfalls integriert werden sollen.

Peer-to-peer-Netzwerke zwischen Unternehmen und öffentlichen Verwaltungen, die über diese Netzwerke Daten austauschen oder Verwaltungsaufgaben abwickeln, stellen eine weitere, ebenfalls kommerzielle Anwendung dar. Hierzu zählt insbesondere der Bereich des E-Government, wie De Santis, Scannapieco und Catarici in [SSC03] bei der Erläuterung des DaQuinCIS-Framework vorstellen. Dabei findet ein Ad-hoc-Zusammenschluss mit dezentralisierter Kontrolle statt.

Abschließend sei noch auf Ostländer hingewiesen, die in [Os01] darauf aufmerksam macht, dass im Bereich der Geowissenschaften eine solide Datengrundlage von hoher Bedeutung ist. Konkret weist sie im Kontext der Untersuchung des Stoffaustrags ausgewählter Einzugsgebiete im Bergischen Land darauf hin, dass zur Auswahl relevanter Qualitätskriterien für diesen Bereich zwei Normen der International Standard Organization (ISO) entwickelt werden. In ISO 19113 werden die Qualitätsdimensionen beschrieben, in ISO 19114 die Art und Weise, wie die dazu gewonnenen Informationen zu bewerten sind.

2.4 Zusammenfassende Betrachtung der Anwendungsdomänen

Allen Anwendungen gemeinsam ist, dass Daten der Domäne in heterogenen und verteilten Datenquellen vorliegen und abgefragt werden. Damit stellt sich die

Aufgabe, die Daten der Quellen zu bewerten, um die für den Nutzer besten Datenquellen herauszusuchen und Anfragen an diese zu stellen, die Daten zusammenzuführen und dem Nutzer zu präsentieren. An dieser Stelle können wir den Aspekt der unterschiedlichen Gewichtung nochmals herausgreifen und zwar in der Hinsicht, dass die Bewertung von Qualitätsmerkmalen in verschiedenen Anwendungsdomänen differiert. In biologischen Datenbanken sind teilweise Antwortzeiten von mehreren Stunden üblich, während ein Nutzer, der eine Internetsuchmaschine abfragt, maximal einige Sekunden warten wird; dieser Wert wiederum unterscheidet sich von Nutzer zu Nutzer.

Bei der Betrachtung der Anwendungsdomänen fällt bei den Suchmaschinen für das Internet auf, dass inzwischen Google den Markt deutlich beherrscht; nach [We06] beträgt der Anteil von Suchanfragen über Google 86 %. An zweiter Stelle folgt Yahoo mit 3,8 %, die erste Metasuchmaschine ist Meta.ger mit 0,3 % (2001 noch mit 5 %). Damit spielen Metasuchmaschinen für das Internet heute offenbar keine große Rolle mehr. In diesem Bereich hat ein Anbieter alle vom Nutzer gewünschten Daten verfügbar und damit alle Konkurrenten verdrängt. Es wird interessant sein zu beobachten, ob in den anderen Bereichen eine ähnliche Entwicklung stattfindet. Diese könnte in einer etwas anderen Form darin bestehen, dass ein Anbieter die Datenintegration vornimmt und die bereits nach Qualitätskriterien integrierten Daten den Nutzern anbietet. Somit könnte der Nutzer die bereits integrierten Daten an einer Stelle abfragen und müsste nicht mehr selbst für die Datenintegration sorgen. Dies wäre eine Dienstleistung, die darauf beruht, dass Information ein Produkt von signifikantem und immer weiter steigendem Wert ist.

2.5 Architekturmodell

Wir stellen zur Betrachtung der Datenqualität in Anlehnung an [Na02] ein allgemeines Architekturmodell für heterogene Informationssysteme und darauf aufbauend die Mediator-Wrapper-Architektur vor.

Grundsätzlich kann man ein heterogenes Informationssystem wie in Abbildung 1 darstellen. Der Nutzer stellt Anfragen an das Informationssystem, das wiederum verschiedene Datenquellen abfragt und dem Nutzer ein Ergebnis präsentiert. Im Anfrageprozess findet nur eine Interaktion des Nutzers mit dem Informationssystem statt.

Wir gehen davon aus, dass die Datenquellen verteilt sind und der Zugriff über elektronische Netze, insbesondere das Internet, erfolgt. Die Datenquellen können heterogen sein, d.h., Informationen werden in unterschiedlicher Art und Weise gespeichert oder nach außen dargestellt. So können Daten als einfache Dateien, als Werte einer HTML-Seite oder in irgendeiner Art von Datenbanksystem gespeichert sein und entsprechend exportiert werden. Diese Aspekte werden in [Na02] als technische und syntaktische Heterogenität gekennzeichnet. In einer Mediator-Wrapper-Architektur ist für jede Datenquelle mindestens ein Wrapper vorhanden, der diese Aspekte der Heterogenität verbirgt und ein relationales Datenbankschema zur Verfügung stellt.

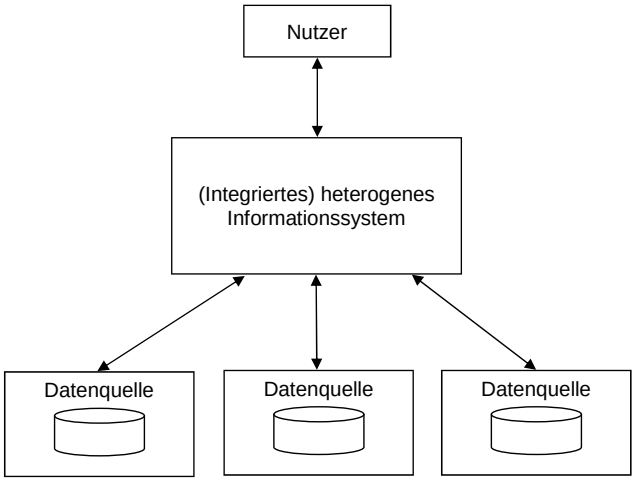

Abbildung 1. Allgemeines Architekturmodell heterogener Informationssysteme

In der Mediator-Wrapper-Architektur besteht der Kern des Informationssystems aus einem Mediator, der sich dem Nutzer wiederum als ein relationales Datenbankschema darstellt. Der Nutzer stellt Anfragen gegen dieses Schema des Mediators. Der Mediator nutzt die Wrapper, um die Datenquellen abzufragen und kombiniert deren Ergebnisse zu einem Gesamtergebnis. Die Schemata der Wrapper stellen meistens nicht die gleiche Attributmenge bereit, wie sie im Schema des Mediators vorhanden ist, da die Datenquellen unterschiedliche Aspekte der Daten vorhalten. Außerdem ist es möglich, dass Datenquellen mehrere, unter Umständen sogar unterschiedliche Werte für das gleiche Tupel exportieren. Diese Aspekte, die Naumann als semantische Heterogenität bezeichnet, müssen beim Mediator bearbeitet und vor dem Nutzer verborgen werden.

Die Mediator-Wrapper-Architektur ist in Abbildung 2 dargestellt, wobei beim Nutzer das integrierte Schema dargestellt wird. Die Informationen für die Spalte A werden im Beispiel der Abbildung 2 aus zwei Wrappern abgefragt und die Datenquelle 2 stellt zwei Wrapper bereit.

Eine solche Architektur, bei der beim heterogenen Informationssystem ein Gesamtschema vorliegt, wird als local-as-view bezeichnet. Alle weiteren Betrachtungen beziehen sich auf den Mediator. Wir nehmen das integrierte Schema beim Mediator sowie die Wrapper als gegeben an. Das integrierte Schema bezeichnen wir wie Naumann als *Universal Relation*. Wir gehen weiterhin davon aus, dass die Universal Relation einen Primärschlüssel hat, dass gleiche Attributnamen sich auf gleiche Eigenschaften beziehen[2] und es für eine beliebige Teilmenge von

[2] Dies bezeichnet Naumann in [Na02], Definition 2.2.2, als "Universal Relation Scheme Assumption".

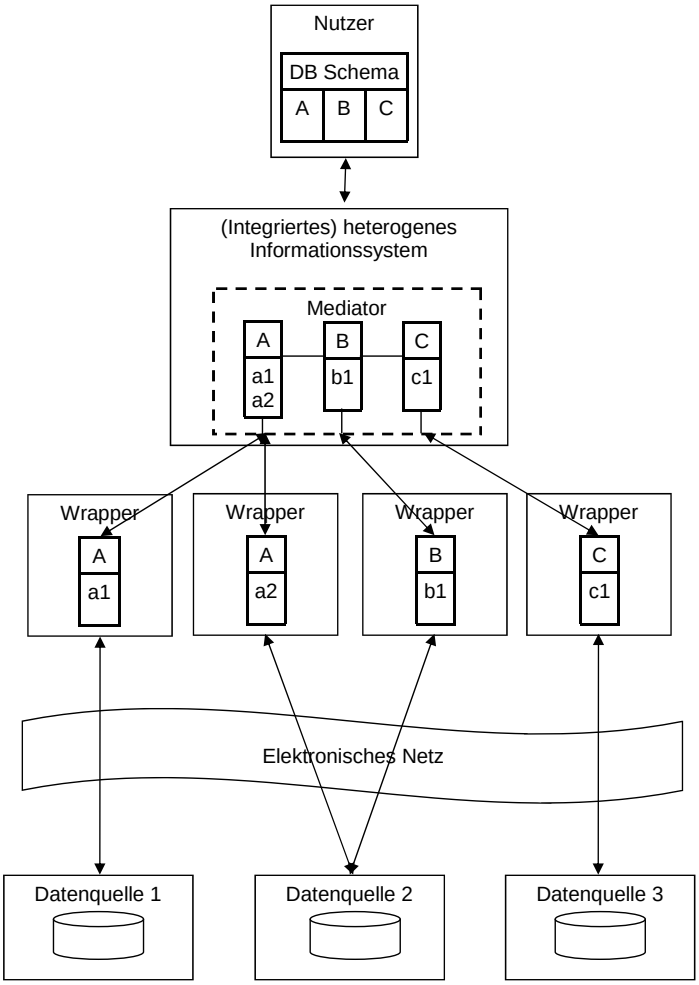

Abbildung 2. Mediator-Wrapper-Architektur

Attributen der Universal Relation eine eindeutig bestimmte Relation gibt, d.h., Attribute hängen nicht zyklisch voneinander ab[3].

3 Qualitätskriterien

Mit der im vorherigen Abschnitt skizzierten Architektur beschreiben wir Qualitätskriterien für Daten. Dabei unterscheiden wir inhaltsbezogene, technische, intellektuelle und präsentationsbezogene Kriterien und stellen eine Auswahl aus den von Naumann in [Nau99] und [Nau02] identifizierten Kriterien vor. Für die Übersetzung ins Deutsche nutzen wir bis auf wenige Ausnahmen die Begriffe aus [NR00]. Abschließend zeigen wir beispielhaft die Auswahl von Qualitätskriterien für zwei konkrete Anwendungsdomänen auf.

3.1 Inhaltsbezogene Qualitätskriterien

Inhaltsbezogene Kriterien beziehen sich auf Inhalte einer Datenquelle, d.h., sie beschreiben Eigenschaften der Daten, die in einer Datenquelle gespeichert sind und die diese an den Mediator und letztlich an den Nutzer liefert.

Genauigkeit beschreibt den Anteil von Daten, die frei von Datenfehlern sind. Unter Datenfehlern fassen wir alle den Daten an sich zuordenbaren Fehler zusammen. Dazu gehören doppelte Primärschlüssel, Schlüssel außerhalb zulässiger Wertebereiche, nicht erlaubte Zeichen und ähnliche Fehler. Wir drücken die Genauigkeit durch den Quotient der Anzahl der korrekten Datensätze in der Datenquelle und aller Datensätze der Quelle aus, so dass der Wert im Intervall [0...1] liegt.

Im Anwendungskontext von Börsenkursen lässt sich die Wichtigkeit verdeutlichen: Gibt es zu einer Aktie mehrere Einträge in einem Kursinformationssystem, so kann der Mediator der Aktie keinen dieser Werte eindeutig zuordnen bzw. der Nutzer daraus keinerlei Information ziehen, da er nicht weiß, nach welchem Wert er sich richten soll, ob er kaufen oder verkaufen soll.

Es sei darauf hingewiesen, dass es hier nicht um die im Anwendungskontext fachliche Korrektheit geht, wie z.B., ob eine gefundene Internet-Seite zu den Suchwörtern passt; dies wird unter Relevanz aufgegriffen.

Vollständigkeit bezeichnet das Verhältnis von Not-Null-Werten zu allen Werten in der Anwendungsdomäne, wie sie durch die Universal Relation erfasst ist. Die Anzahl der Not-Null-Werte ist die Anzahl von Attribut-Wert-Paaren, die in einer Datenquelle gespeichert sind. Da wir nicht wissen, wie viele Werte innerhalb der Anwendungsdomäne vorhanden sind, schätzen wir diese Anzahl durch die maximal verfügbaren Werte bei Betrachtung des Datenbestands aller Quellen ab.

Die Zahl der maximal verfügbaren Werte ist die Anzahl der Attribute der Universal Relation multipliziert mit der Anzahl an verfügbaren Tupeln, d.h. Tupeln, die in der Universal Relation verschiedene Einträge für den Primärschlüssel

[3] In Definition 2.2.3 in [Na02] als "Unique Role Assumption" zu finden.

haben. Zur Verdeutlichung können wir uns alle Werte in der Universal Relation materialisiert vorstellen. Dann multiplizieren wir die Anzahl der Zeilen mit der Anzahl der Spalten.

Der Wert für die Verfügbarkeit liegt mit dieser Definition ebenfalls im Intervall [0...1].

Wenn wir im Beispiel des Börsenszenarios als Attribute in der Universal Relation den aktuellen Preis, die Preisentwicklung und ein Firmenprofil modellieren, dann muss eine vollständige Datenquelle (Vollständigkeit = 1) für alle Aktien, zu denen es Werte in Informationssystemen gibt, alle drei Informationen vorhalten. Im Allgemeinen werden für eine Vollständigkeit, die gegen 1 geht, mehrere Datenquellen benötigt, wie z.B. die Informationssysteme aller Börsen.

Interpretierbarkeit beschreibt, in wie weit die gelieferten Informationen den fachlichen Anforderungen des Nutzers genügen. Darunter fallen die Sprachen, die der Nutzer versteht, inklusive Fachsprachen oder Einheiten, die ihm bekannt sind. Damit ist die Notwendigkeit eingeschlossen, dass bei einer Datenquelle alle Informationen ausreichend erklärt, definiert und dokumentiert sind.

Dieses Kriterium ist im Gegensatz zu den ersten beiden subjektiv und differiert bei verschiedenen Nutzern. So muss der Nutzer eines Börseninformationssystems wissen, in welcher Währung die Wertangaben von Aktien sind[4].

Relevanz bezeichnet, wie stark die Daten den Bedürfnissen und Anforderungen des Nutzers genügen. Dieser Wert ist subjektiv für jeden Nutzer und darüber hinaus unterschiedlich für verschiedene Datensätze. Naumann schränkt diesen Begriff in [Na02] so weit ein, dass er davon ausgeht, dass jedes Resultat, das korrekt im Bezug auf eine Nutzeranfrage ist, auch relevant ist. Andernfalls war entweder die Anfrage falsch für die Informationen, die der der Nutzer wollte oder nicht ausreichend spezifiziert. Damit verschiebt Naumann die Verantwortung ein wenig auf den Nutzer, nämlich, dass dieser korrekte und eindeutig spezifizierte Suchanfragen stellt und darauf, dass die Suchanfragen korrekt bearbeitet werden, ohne für diese korrekte Bearbeitung eine Definition anzugeben. Eine differenzierte Betrachtung, die beide Aspekte, den Nutzer und das Informationssystem, berücksichtigt, wäre an dieser Stelle aus unserer Sicht zu empfehlen.

Um im Beispiel des Börseninformationssystems zu bleiben, ist für einen Nutzer, der sich über den Verlauf der Aktie A informieren will, der aktuelle Kurs nicht relevant. Für einen anderen Nutzer, der bei einem bestimmten Kurs in die Aktie investieren will, ist dagegen nur der aktuelle Kurs relevant.

Die Relevanz ist im Bereich von Internetsuchmaschinen (bei uns implizit immer Metasuchmaschinen) von großer Bedeutung, da auf eine Suchanfrage nur solche Webseiten angezeigt werden sollten, die im Zusammenhang mit der Anfrage stehen. Auf Grund von Synonymen, Homonymen, Einzahl-, Mehrzahl-Problemen und Ähnlichem ist die Relevanz in dieser Anwendung oftmals nicht automatisiert erfassbar, da durch diese Effekte natürlicher Sprachen Unschärfen auftreten. Dies ist nach Dessloch in [De03] mit dem Begriff Precision erfasst,

[4] In unserer Standardarchitektur tritt dieses Problem nicht auf, da es an der Schnittstelle zwischen Mediator und Wrapper gelöst wird. Der Mediator integriert Daten in der Währung, die die Universal Relation vorgibt.

wobei der Precision-Wert den Anteil der relevanten Dokumente im Suchergebnis beschreibt.

3.2 Technische Qualitätskriterien

Technische Qualitätskriterien beschreiben alle Aspekte, die durch Hard- und Software der Datenquellen und des Informationssystems, bestehend aus dem Mediator und den Wrappern, bestimmt sind. Dazu gehören die Verbindungen des Informationssystems zu den Datenquellen. Damit sind neben Aspekten die Datenquellen betreffend, alle technischen Aspekte der integrierenden Einheiten und des Netzwerkes erfasst.

Verfügbarkeit beschreibt den Anteil an Anfragen, auf die das System innerhalb einer bestimmten Zeitspanne ein Ergebnis liefern kann. Damit ist ein Wert im Intervall [0...1] beschrieben, der angibt, mit welcher statistischen Wahrscheinlichkeit eine Anfrage zu einem Ergebnis führt.

Da wir die Datenquellen als autonom annehmen, haben wir auf ihre Verfügbarkeit keinen Einfluss. Die Verfügbarkeit hängt auch von der Zuverlässigkeit des Netzes ab, die oftmals im Verlauf des Tages schwankt, was durch längerfristige (wöchentliche, monatliche oder jährliche) Belastungsmuster überlagert und verstärkt werden kann. Je geringer die Verlässlichkeit der beteiligten Quellen ist, desto mehr Datenquellen müssen wir abfragen. Dies muss unter Umständen dynamisch während der Abfrage entschieden werden, was die Festlegung eines vorher definierten Ablaufplans erschwert.

Beim Beispiel der Börsenabfrage haben wir bereits darauf verwiesen, dass es für Informationen über Aktien, die an unterschiedlichen Börsen gehandelt werden, erforderlich sein wird, mehrere Datenquellen abzufragen, um alle Informationen zu erhalten. Fällt in diesem Szenario die Anbindung an das Informationssystem einer Börse aus, können wir die fehlenden Informationen vielleicht bei großen Bankinstituten abfragen und damit den Ausfall der primären Quelle durch eine sekundäre kompensieren.

Genau wie die folgenden Kriterien Latenzzeit, Antwortzeit, Aktualität und Preis ist die Verfügbarkeit ein Kriterium, das oftmals nicht vor Bearbeitung der Anfrage festgestellt werden kann, aber dennoch objektiv feststellbar ist, da es mit einfachen Methoden messbar ist.

Latenzzeit misst die Zeitspanne, die es dauert, bis die ersten Antworten, also die ersten Daten eintreffen. Diese Zeit wird im Allgemeinen in Sekunden angegeben. Alle Aspekte, die bei der Verfügbarkeit bereits genannt wurden, spielen hierbei eine Rolle und können allgemein mit dem Workload der Datenquelle umschrieben werden. Bei umfangreichen Ergebnissen kann die Latenz wichtig sein, wenn bereits die ersten Daten verarbeitet werden können, während weitere Daten ankommen.

Bei Metasuchmaschinen im Internet kann die Wichtigkeit dieses Wertes verdeutlicht werden. Im Allgemeinen genügen dem Nutzer die ersten 10 Ergebnisse, selbst wenn sehr viele Treffer ermittelt wurden. Die ersten Ergebnisse sollten dargestellt werden, sobald sie verfügbar sind. Von Suchmaschinen werden im

Allgemeinen nur die ersten 10 bis maximal 100 Ergebnisse übermittelt und nicht das Gesamtergebnis.

Antwortzeit beschreibt im Unterscheid zur Latenzzeit die Zeitspanne, bis das gesamte Ergebnis übermittelt wurde. In Ergänzung der Einflussfaktoren, die von der Latenzzeit bekannt sind, kann die Komplexität der Anfrage eine entscheidende Rolle spielen.

Hier verweisen wir nochmals auf das bei der Latenzzeit vorgestellte Szenario einer Metasuchmaschine. Der Nutzer ist meistens gar nicht am gesamten Ergebnis interessiert, so dass die Antwortzeit keine Rolle spielt (von Ausnahmefällen abgesehen).

Aktualität bezeichnet das durchschnittliche Alter der Daten. Dies ist ein Qualitätskriterium, das in verschiedenen Anwendungsdomänen unterschiedlich aufgefasst werden kann. Bei Internetsuchmaschinen wird man darunter die Zeit verstehen, in der die Suchmaschine die Seite zuletzt besucht und indexiert hat. Dies wird sich im Bereich von Tagen bewegen; ist das durchschnittliche Alter zu hoch, sind zu viele schon entfernte oder geänderte Seite zu erwarten; genauso werden neue Seiten, die noch nicht im Suchindex geführt werden, aber wichtige Informationen und Links liefern können, nicht gefunden.

Für Börsenkurse ist die Aktualität entscheidend und muss im Zeitbereich von Sekunden angeben werden.

Preis ist ein Qualitätskriterium, das immer wichtiger wird, je mehr es einen Markt für kommerzielle Datenanbieter gibt. Wichtig ist neben der Bewertung des Preises in einer einheitlichen Währung die Berücksichtigung des Preismodells. Zu unterscheiden sind Preismodelle, bei denen auf Basis von Abonnements eine unbegrenzte oder fest vorgegebene Nutzung der Datenquelle möglich ist, und Modelle, bei denen für jede Anfrage ein bestimmter Betrag fällig wird. Vorstellbar ist auch ein variabler Preis je nach Art und Komplexität der Anfrage.

Sicherheit fasst alle Methoden zusammen, die für eine sichere Übertragung aller Eingaben und Informationen des Nutzers zur Datenquelle und zurück eingesetzt werden. Dazu zählen kryptographische Algorithmen, die verwendete Netzarchitektur, Login-Mechanismen, Anonymität der Datenverarbeitung, Sicherung der Server gegen Diebstahl persönlicher Daten, sicheres Bezahlen und vieles mehr.

Für Internetsuchmaschinen ist dies kein kritisches Kriterium, für Börseninformationssysteme, insbesondere solche, die die Möglichkeit zum Kauf anbieten, dagegen schon.

3.3 Intellektuelle Qualitätskriterien

Mit intellektuellen Kriterien untersuchen wir subjektive Einstellungen und Meinungen über Datenquellen. Diese sind daher alle durch den Nutzer auf Skalen zu beurteilen und können nicht maschinell erfasst werden.

Reputation beschreibt, ob die Datenquelle einen "guten Ruf" hat. Viele Benutzer ziehen Daten vor, die aus einer allgemein anerkannten Quelle stammen, genauer gesagt aus einer Quelle, die sie für allgemein anerkannt halten, die in ihren Augen eine hohe Reputation hat.

In [Na02] berichtet Naumann, dass insbesondere in biologischen Datenbanken Forscher Daten von Instituten bevorzugen, die ein großes Ansehen in der Fachwelt haben. Darüber hinaus wird den Ergebnissen des eigenen Forschungsinstituts mehr vertraut als externen Daten.

Objektivität ist der Grad, zu dem die Daten unverfälscht und unbeeinflusst geliefert werden.

Suchmaschinen bieten kommerziellen Anbietern die Möglichkeit, höher bewertet zu werden, als dies durch die normalen Algorithmen geschieht, um Kunden auf die Seite des Unternehmens zu lenken. Damit verfälschen und beeinflussen sie die Daten. Der Benutzer ist in diesem Beispiel machtlos, da er solche Manipulationen nicht erkennen kann, sofern die Links nicht als Werbeanzeige gekennzeichnet sind.

3.4 Präsentationsbezogene Qualitätskriterien

Zu den präsentationsbezogenen Qualitätskriterien zählen Merkmale eines konkret gelieferten Datensatzes. Viele Aspekte werden dabei durch die Mediator-Wrapper-Architektur verdeckt, da wir davon ausgehen, dass als Ergebnis stets Tupel eines relationalen Schemas dem Nutzer geliefert und präsentiert werden. In dieser Architektur werden Daten immer im selben Format präsentiert (*Konsistente Darstellung*) und die Datendarstellung hängt nicht mehr von der Darstellung der Datenquelle ab, sondern von der Universal Relation. Somit ist die Passung von Daten und Darstellung (*Einfachheit der Darstellung*) eine Aufgabe des Mediators, der damit die *Verständlichkeit der Darstellung* determiniert. Über diese knappe Darstellung hinaus betrachten wir im Folgenden zwei Aspekte genauer.

Datenmenge ist die Größe des gelieferten Suchergebnisses, die sich in Byte ausdrücken lässt. Für Anwendungen, bei denen es festgelegte Datensätze oder Einheiten des Datentransfers gibt, wie z.B. Links in Suchmaschinen, sollte die Datenmenge in diesen Einheiten erfasst werden. Die Möglichkeit der automatisierten Messung bleibt davon unberührt.

Prüfbarkeit ist dann gegeben, wenn die Daten von einer weiteren, unabhängigen Quelle bestätigt werden können. Dies können bei biologischen Forschungsergebnisse Resultate anderer Institute sein, bei Nachrichtenmeldungen der Nachrichtendienst, der die Meldung veröffentlich hat oder bei einem Suchergebnis im Internet die referenzierte Internetseite. Ein Suchergebnis einer Suchmaschine lässt sich somit direkt und einfach durch Besuch der gefundenen Seite verifizieren, während bei komplexen und spezifischen Daten, wie Forschungsergebnissen, oftmals keine Möglichkeit einer unabhängigen Überprüfung besteht. Gibt die Datenquelle keine Informationen über den Ursprung der Daten an, ist also die primäre Quelle nicht bekannt, wird die Überprüfung zusätzlich erschwert[5].

[5] Mit den rechtlichen Aspekten einer Veröffentlichung ohne Quellenangabe befassen wir uns hier selbstverständlich nicht. Möglich wäre aber auch, dass die Herkunft in der Universal Relation nicht modelliert wurde.

3.5 Auswahl von Qualitätskriterien

Bereits bei der Vorstellung der einzelnen Qualitätskriterien haben wir auf verschiedene Anwendungsbereiche hingewiesen und die unterschiedliche Wichtigkeit deutlich gemacht. Es ist nicht für jede Anwendungsdomäne möglich, alle diese Qualitätskriterien zu erheben[6]. Es sollte vielmehr eine sinnvolle Auswahl für die konkrete Anwendung getroffen werden. Dabei bedürfen einige Aspekte einer anwendungsspezifischen Interpretation. Wir zeigen im Folgenden für eine Suchmaschine in Tabelle 1 und ein Börseninformationssystem in Tabelle 2 beispielhaft eine mögliche Auswahl mit passenden Interpretationen auf. Die Auswahl der Kriterien hängt von vielen Faktoren ab. Wir nehmen die Auswahl in Anlehnung an [NR00] vor, weichen aber bei einigen Punkten leicht ab.

Tabelle 1. Qualitätskriterien für eine Suchmaschine

Relevanz	Nur Seiten, die im Zusammenhang mit der Suchanfrage stehen, sind im Ergebnis aufgelistet
Genauigkeit	Qualität der Resultats-Sortierung
Aktualität	Update-Frequenz des Suchindexes in Tagen
Verfügbarkeit	Prozent der Zeit, in der die Suchmaschine erreichbar ist
Vollständigkeit	Prozentualer Anteil der Webseiten, die im Suchkatalog sind
Latenzzeit	Sekunden, bis der erste Link im Browser dargestellt wird

Tabelle 2. Qualitätskriterien für ein Börseninformationssystem

Verfügbarkeit	Prozent der Zeit, in der das System erreichbar ist
Vollständigkeit	Prozent der gelisteten Aktien, die an der Börse im DAX gehandelt werden
Objektivität	Unabhängigkeit der Quelle
Preis	Geld pro Anfrage
Antwortzeit	Zeit, bis alle Börsenkurse übertragen wurden
Datensicherheit	Verschlüsselung des Datentransfers und Sicherheit des Servers
Aktualität	Keine künstliche Verzögerung von Kursänderungen durch den Anbieter

4 Erhebung von Qualitätsdaten

Nach der Identifikation von Qualitätskriterien stellt sich das Problem, diese Informationen tatsächlich zu erheben. Für die klassischen Kriterien in zentralisier-

[6] Zumal wir hier schon eine Auswahl aus einer Fülle weiterer Kriterien getroffen haben.

ten Datenbanksystemen, Antwortzeit und Durchsatz, können Messungen mit Hilfe von Benchmarks durchgeführt werden. Für viele der vorgestellten Qualitätskriterien ist eine vollständig automatisierte Datenerhebung nicht möglich, da sie subjektiv sind.

Wesentlich für eine gute Erhebung ist eine möglichst genaue Definition. Bei allen Werten, die keine natürlichen Einheiten wie Sekunde haben, sondern wie die Reputation auf festgelegten Skalen verankert werden müssen, ist eine gute Gestaltung und verständliche Erklärung dieser Skalen notwendig. Dazu bietet es sich an, Referenzpunkte festzulegen, um Nutzern Anhaltspunkte für die Bewertung zu geben.

4.1 Klassen und Quellen von Qualitätsdaten

Für einige der nicht messbaren Kriterien wären die Datenquellen in der Lage Informationen zur Verfügung zu stellen, z.B. über die Aktualität oder die Objektivität. Dies erfolgt aber in der Regel nicht, obwohl es Modelle und Vorschläge für Standards für Qualitätsdaten autonomer Datenquellen gibt. Ein Wettbewerb zwischen Anbietern könnte zu der Veröffentlichung von unabhängigen Qualitätsdaten nach einheitlichen Modellen führen und diese Situation verbessern.

Bei anderen Kriterien, wie der Anzahl der Not-Null-Werte, wäre es prinzipiell möglich eine exakte Messung durchzuführen, aber für große Datenbestände es ist praktisch unmöglich, den gesamten Datenbestand einzulesen und zu untersuchen. Daher müssen in solchen Fällen Urteile auf der Basis von Stichproben erfolgen.

Hat man diese Daten erhoben, stellt sich im nächsten Schritt das Problem, dass alle Datenquellen unabhängig sind und somit keiner Kontrolle des Informationssystems unterliegen. Regelmäßige Stichproben sind erforderlich, um Qualitätsänderungen festzustellen. Ansonsten besteht die Gefahr, dass durch alte Qualitätswerte eine Datenquelle unter- oder überschätzt wird.

Wir unterscheiden in diesem Kapitel die Qualitätskriterien nach subjektiven, objektiven und anfragespezifischen Kriterien. Subjektive Qualitätskriterien können nur vom Nutzer bewertet werden. Dazu zählt z.B. die Reputation. Dagegen können objektive Qualitätskriterien aus einer Datenanalyse gewonnen werden, wie dies beispielsweise bei der Messung der Anzahl von Not-Null-Werten möglich ist. Die anfragespezifischen Qualitätskriterien verändern sich von Anfrage zu Anfrage, wie die Latenzzeit, die unter anderem von der Tageszeit und der Komplexität der Anfrage abhängt. Die genannten Zusammenhänge zwischen Klassen von Qualitätskriterien und ihren Quellen zur Erhebung sind in Abbildung 3 graphisch dargestellt.

Da wir in diesem Kapitel die Qualitätskriterien nach diesen drei Klassen untersuchen, ist die Einordnung der in Kapitel 3 vorgestellten Qualitätskriterien in diese Klassen in Tabelle 3 dargestellt.

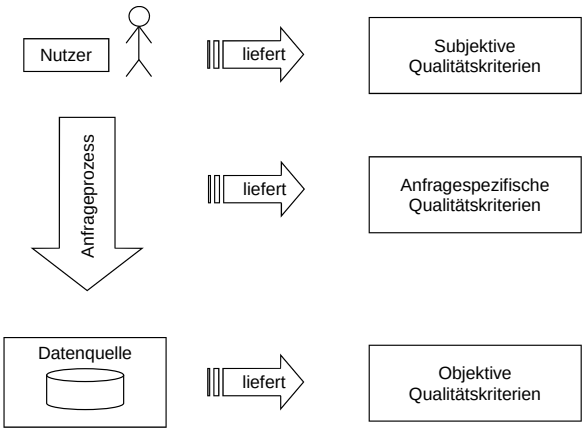

Abbildung 3. Klassen und Quellen von Qualitätskriterien

Tabelle 3. Klassen von Qualitätskriterien

Subjektive Qualitätskriterien	Interpretierbarkeit Relevanz Reputation Einfachheit der Darstellung Verständlichkeit der Darstellung
Objektive Qualitätskriterien	Vollständigkeit Aktualität Preis Sicherheit Objektivität Prüfbarkeit
Anfragespezifische Qualitätskriterien	Genauigkeit Verfügbarkeit Latenzzeit Antwortzeit Konsistente Darstellung Datenmenge

4.2 Erhebung subjektiver Qualitätskriterien

Subjektive Qualitätskriterien müssen auf vorher festgelegten Skalen bewertet werden. Die Skalen sollten einfach sein, wie z.B. ein Punktesystem von 10 (sehr gut) bis 0 (sehr schlecht). Eine Bewertung hat immer nur Gültigkeit für den Benutzer, der sie vorgenommen hat. Der Nutzer muss für eine Bewertung Zeit investieren, damit sie möglichst genau ist. Allerdings wollen Nutzer das System nutzen, ohne lange Zeit Bewertungen eingeben zu müssen. Daher sollten diese Werte mit einfachen Fragenbögen erfasst werden, die die Festlegungen auf den Skalen erlauben. Für jeden Nutzer muss ein entsprechendes persönliches Profil erstellt und gespeichert werden.

Im System können Standard-Werte vorgegeben sein, die sich entweder aus dem Durchschnitt aller Nutzer ergeben oder von einem Experten, z.B. dem Systemadministrator vorgegeben werden. Außerdem können zur Erklärung Referenzpunkte mit besonders schlechten oder besonders guten Beispielen angegeben werden. Darüber hinaus muss es möglich sein, diese Bewertung jederzeit zu ändern. Stellt das System bei einer Stichprobe eine deutliche Qualitätsänderung fest, sollte es eine Änderung der Bewertung vorschlagen.

Bei diesem Vorgang ist zu unterscheiden zwischen Kriterien, die sich auf den in der Datenquelle gespeicherten Datenbestand oder ein konkretes Ergebnis beziehen. Während ersteres an Hand einer Stichprobe untersucht und bewertet werden kann, muss bei konkreten Ergebnissen prinzipiell nach jeder Anfrage eine neue Bewertung vorgenommen werden, wie z.B. bei der Relevanz, die sich auf das Ergebnis einer Anfrage bezieht.

4.3 Erhebung anfragespezifischer Qualitätskriterien

Bei den anfragespezifischen Qualitätskriterien wird eine einmalige Suchanfrage bewertet, wobei alle Kriterien in dieser Klasse eine genau festgelegte Einheit haben. Die Bewertung geschieht mit der Absicht, aus den bisherigen Erfahrungen Prognosen über zukünftige Anfragen treffen zu können. Die Bewertung kann von vielen Einflussgrößen abhängen, wie der Tageszeit und der Komplexität der Anfrage. Somit sind die Werte in dem Moment der Suchanfrage exakt, verlieren aber über die Zeit ihre Gültigkeit oder werden zumindest unschärfer. Da alle Kriterien automatisch erfasst werden können, sollten diese Daten laufend gesammelt werden, um Tendenzen zu erkennen und in die Bewertung mit einfließen zu lassen. Allerdings sollte die Analyse die Anfrageverarbeitung möglichst nicht verzögern. Es ist wenig sinnvoll, den Nutzer langen Antwortzeiten auszusetzen, die von zusätzlichen Datenanalysen verursacht werden.

4.4 Erhebung objektiver Qualitätskriterien

Objektive Qualitätskriterien können zum größten Teil automatisch erhoben werden. Dazu ist eine Abfrage der Datenquelle notwendig, um Werte zur Verarbeitung zu haben. Bei größeren Datenquellen sind gute Verfahren zur Stichprobenbildung und Hochrechnung zu verwenden. Automatisierte Bewertungen sollten

möglichst regelmäßig durchgeführt werden und daher einfache Verfahren nutzen, um das Informationssystem nicht zu stark zu belasten. Damit können z.B. die Vollständigkeit überprüft oder zumindest gut geschätzt werden oder die Sicherheit, in dem die zur Übertragung verwendeten Protokolle eingelesen werden. Für alle diese Maßnahmen werden Algorithmen aus dem Bereich der Parser eingesetzt.

Eine Sonderstellung nehmen der Preis und die Objektivität ein. Der Preis ergibt sich aus Verträgen und muss manuell eingegeben werden. Bei der Objektivität ist Expertenwissen notwendig, um diese zu beurteilen. Für die Bewertung durch einen Experten können wiederum Methoden aus dem Bereich der subjektiven Qualitätskriterien eingesetzt werden. Allerdings sollte sich der Experte um ein möglichst objektives Urteil bemühen und nicht seine eigene Meinung in den Vordergrund stellen.

5 Gewichtung von Qualitätskriterien

Nachdem wir im vorherigen Kapitel 4 darauf eingegangen sind, wie einzelne Qualitätskriterien ermittelt und bewertet werden können, stellt sich nun die Frage, wie daraus ein Gesamtergebnis für eine Datenquelle berechnet werden kann. Denn nur wenn das möglich ist, kann eine Auswahl auf Basis der ausgewählten Qualitätskriterien vorgenommen werden. Dabei muss, wie schon mehrfach erwähnt, berücksichtigt werden, dass nicht alle der ausgewählten Qualitätskriterien gleich wichtig sind. Die Einschätzung, welche Kriterien besonders wichtig sind, kann sich von Nutzer zu Nutzer unterscheiden. Diese Probleme wollen wir in diesem Kapitel diskutieren, indem wir einerseits präzise mathematische Definitionen einführen und diese anderseits, wo immer möglich, an einem durchgängigen Beispiel erläutern.

5.1 Qualitätsmodell für Datenquellen

Wir stellen ein Qualitätsmodell nach [Na02] vor, das eine mathematische Modellierung erlaubt. Dazu bezeichnen S_i die verschiedenen Datenquellen, die in einem heterogenen Informationssystem integriert werden. Wir gehen davon aus, dass wir j geeignete Qualitätskriterien für die Beurteilung des Informationssystems ausgewählt haben.

Definition 1. *Sei S_i eine von insgesamt n Datenquellen ($i = 1, ..., n$) und seien d_{ik} die Bewertungen der j Qualitätskriterien, also $k = 1, ..., j$. Dann heißt der Vektor $QV(S_i) = (d_{i1}, ..., d_{ij})$ der Qualitätsvektor der Datenquelle S_i.*

Beispiel 1. Wir betrachten eine Metasuchmaschine M, die zwei Suchmaschinen S_1 und S_2 abfragt. Wir beurteilen die beiden Suchmaschinen nach den Kriterien Relevanz und Aktualität. S_1 liefert im Ergebnis immer 50 % zutreffende Links und aktualisiert ihren Suchindex alle 3 Tage. S_2 liefert im Ergebnis immer 75 %

zutreffende Links und aktualisiert ihren Suchindex alle 10 Tage. Somit sind die Qualitätsvektoren[7]:

$$QV(S_1) = (0.5, 3)$$

$$QV(S_2) = (0.75, 10)$$

Nun stellt sich die Frage, wie wir solche Vektoren vergleichen können, wie wir eine Ordnung auf dem j-dimensionalen Raum der Qualitätsvektoren festlegen können. In Beispiel 1 sehen wir, dass die Wertebereiche unterschiedlich sind. Dazu kann das Problem kommen, dass die Werte nicht gleichmäßig über den Bereich gestreut sind. So liegt die Verfügbarkeit im Allgemeinen nicht im Bereich [0...1], sondern im Bereich $[0,95...0,999999]$. Die Qualitätsvektoren müssen passend skaliert werden.

Im Beispiel ist S_1 aktueller als S_2, liefert dafür aber schlechtere Resultate. Hier ist zu entscheiden, was dem Nutzer wichtiger ist, die Einträge der Qualitätsvektoren müssen gemäß Nutzervorgaben gewichtet werden. Diesen beiden Problemen stellen wir uns in den folgenden Abschnitten.

5.2 Skalierung von Datenqualitätswerten

Wir stellen zwei Möglichkeiten vor, Datenqualitätswerte unabhängig von ihrer tatsächlichen Einheit auf das Intervall [0...1] abzubilden und somit beliebige Werte vergleichen zu können. Eine einfache Methode ist eine Normalisierung über alle Datenquellen, indem wir aus dem Wert d_{ik} einen neuen Wert v_{ik} berechnen durch:

$$v_{ik} = \frac{d_{ik}}{\sqrt{\sum_{i=1}^{n} d_{ik}^2}} \tag{1}$$

In (1) nutzen wir zur Normalisierung die Werte für das betrachtete Qualitätskriterium k in allen Datenquellen $S_1, ..., S_n$. Wir berechnen die euklidische Norm für den Vektor, der alle n angenommenen Werte eines Kriteriums beinhaltet. Somit projiziert Gleichung (1) die Werte in den Bereich [0...1], aber ohne eine gleichmäßige Verteilung zu erreichen. Dafür bleiben proportionale Beziehungen erhalten, wie wir in Beispiel 2 nachvollziehen.

Beispiel 2. Wir berechnen zunächst die euklidischen Normen:

$$\sqrt{\sum_{i=1}^{2} d_{i1}^2} = \sqrt{0,5^2 + 0,75^2} \approx 0,9$$

$$\sqrt{\sum_{i=1}^{2} d_{ij}^2} = \sqrt{3^2 + 10^2} \approx 10,4$$

[7] Der Übersichtlichkeit wegen schreiben wir in Vektoren Kommazahlen immer mit einem Punkt. Damit sind sie leicht von den Kommata zu unterschieden, die die einzelnen Einträge abtrennen.

Daraus erhalten wie die skalierten Qualitätsvektoren:

$$QV(S_1) = (\frac{0.5}{0.9}, \frac{3}{10.4}) \approx (0.6, 0.29)$$

$$QV(S_2) = (\frac{0.75}{0.9}, \frac{10}{10.4}) \approx (0.8, 0.96)$$

Wir sehen, dass sich bei der Relevanz Quelle S_1 und S_2 nach der Skalierung weiterhin im Verhältnis $3 : 4$ befinden. Der Wertebereicht $[0...1]$ wird nicht vollständig genutzt, 0 und 1 werden nicht erreicht.

Wir wollen im Folgenden zwischen positiven und negativen Qualitätskriterien unterscheiden, was wir direkt am Beispiel erläutern.

Beispiel 3. Bei der Metasuchmaschine M ist das erste Qualitätskriterium (Relevanz) positiv, da wir einen möglichst hohen Anteil an relevanten Dokumenten im Suchergebnis haben wollen und somit eine Maximierung des Wertes anstreben. Das zweite Qualitätskriterium (Aktualität) ist negativ, da wir möglichst aktuelle Suchergebnisse und somit eine Minimierung des Zahlenwertes erreichen wollen.

Diese Eigenschaft drückt die Skalierung mit Gleichung (1) nicht aus. Eine andere Art der Skalierung erlaubt uns die Unterscheidung zwischen positiven und negativen Kriterien. Dazu bezeichnen wir mit d_k^{max} den maximal erreichten Wert des Qualitätskriteriums k und analog mit d_k^{min} den minimal erreichten. Dann berechnet sich die Skalierung durch:

$$v_{ik} = \frac{d_{ik} - d_k^{min}}{d_k^{max} - d_k^{min}} \text{für positive Qualitätskriterien} \tag{2}$$

$$v_{ik} = \frac{d_k^{max} - d_{ik}}{d_k^{max} - d_k^{min}} \text{für negative Qualitätskriterien} \tag{3}$$

Damit erreichen wir eine Verteilung auf das gesamte Intervall. Setzen wir in (2) $d_{ik} = d_k^{max}$, so erhalten wir 1 als Ergebnis, mit $d_{ik} = d_k^{min}$ ist das Ergebnis 0. Es wird der volle Wertebereich ausgenutzt, 0 und 1 werden erreicht. Die übrigen Werte werden durch den Nenner in diesen Bereich gestreut. Bestehende Proportionalitäten bleiben nicht erhalten. Dies zeigen wir in Beispiel 4 auf.

Beispiel 4. Es ergibt sich, dass $d_1^{max} = 0.75$ und $d_1^{min} = 0.5$ ist. Außerdem ist $d_2^{max} = 10$ und $d_2^{min} = 3$. Alle Werte sind in unserem Beispiel maximale oder minimale Werte, so dass sich für unsere skalierten Qualitätsvektoren als Ergebnis ergibt:

$$QV(S_1) = (\frac{0.5 - 0.5}{0.75 - 0.5}, \frac{10 - 3}{10 - 3}) = (0, 1)$$

$$QV(S_2) = (\frac{0.75 - 0.5}{0.75 - 0.5}, \frac{3 - 3}{10 - 3}) = (1, 0)$$

5.3 Gewichtung

Die Notwendigkeit der Gewichtung wurde bereits ausführlich dargestellt. Wir definieren den Begriff eines Gewichtungsvektors und verfolgen das Konzept an unserem Beispiel.

Definition 2. *Seien für eine Anwendung j Qualitätskriterien gegeben. Dann heißt jeder Vektor $(w_1, ..., w_j)$ mit $w_k \geq 0, w_k \in \mathbb{R}$ ein Gewichtungsvektor für unsere Anwendung.*

Die üblicherweise eingesetzten Methoden, um mit einem Gewichtungsvektor eine Entscheidung zu treffen, fordern außerdem, dass $\sum_{k=1}^{j} w_k = 1$ gilt.

Eine konkrete Umsetzung dieses Konzeptes ergibt sich, indem der Nutzer auf einer festgelegten Skala, z.B. von 0 bis 10, angibt, wie wichtig ihm das jeweilige Kriterium ist. Die 0 ist explizit zugelassen, damit der Nutzer einzelne Kriterien deaktivieren kann. Um für diesen einfachen Gewichtungsvektor noch die Bedingung $\sum_{k=1}^{j} w_k = 1$ zu erfüllen, können wir den Vektor $w = (w_1, ..., w_j)$ wie folgt zu einem Vektor $w' = (w_1', ..., w_j')$ normalisieren:

$$w_k' = \frac{w_k}{\sum_{l=1}^{j} w_l} \tag{4}$$

Da für alle w_k' der Nenner $\sum_{l=1}^{j} w_l$ gleich ist und im Zähler sukzessive die einzelnen Summanden $w_1, ..., w_j$ aufaddiert werden, ergibt sich die gewünschte Eigenschaft $\sum_{k=1}^{j} w_k' = 1$.

Beispiel 5. Wir setzen unser Beispiel fort und gehen davon aus, dass für unseren Benutzer die Relevanz sehr wichtig ist und er ihr auf einer Skala von 0 bis 10 den Wert 9 gegeben habt. Für die Aktualität nehmen wir als Wert 3 an. Somit ist der Gewichtungsvektor $(9, 3)$. Der normierte Gewichtungsvektor ist $(\frac{9}{12}, \frac{3}{12}) = (\frac{3}{4}, \frac{1}{4}) = (0.75, 0.25)$.

5.4 Auswertung von Qualitätsvektoren

Wir wollen jetzt das ursprüngliche Problem dieses Kapitels lösen, nämlich einer Datenquelle einen Datenqualitätswert zuordnen. Wir stellen drei Methoden vor, die unterschiedliche Schwerpunkte sowie Stärken und Schwächen haben, die wir jeweils kurz aufzeigen.

Als erstes stellen wir die Methode der *Einfachen Additiven Gewichtung* vor, die aus einem skalierten und gewichteten Qualitätsvektor für eine Datenquelle S_i einen Datenqualitätswert $DQ(S_i)$ in [0...1] berechnet. Im Folgenden beziehen wir uns auf diese Methode unter Nutzung der in der Literatur üblichen Abkürzung "SAW" für "Single Additive Weighting".

Dies ist die einfachste Methode und schließt sich direkt an unsere bisherigen Ergebnisse an. Haben wir einen gemäß (2) und (3) skalierten Qualitätsvektor

mit j Qualitätskriterien $QV(S_i) = (v_{i1}, ..., v_{ij})$ und einen gemäß (4) normierten Gewichtungsvektor $(w_1^{'}, ..., w_j^{'})$ so berechnen wir den Datenqualitätswert durch:

$$DQ(S_i) = \sum_{l=1}^{j} w_l^{'} v_{il} \tag{5}$$

Gleichung (5) beschreibt, dass wir das (kanonische) Skalarprodukt des Gewichtungsvektors mit dem Qualitätsvektor berechnen, indem wir komponentenweise multiplizieren und anschließend summieren. Für eine ideale Datenquelle, die für alle Qualitätswerte den maximalen Wert 1 hat, summieren wir den Gewichtungsvektor auf $(v_{il} = 1)$ und erhalten als Gesamtergebnis nach (4):

$$\sum_{l=1}^{j} w_l^{'} v_{il} = \sum_{l=1}^{j} w_l = 1$$

Der minimale Wert ist entsprechend 0, falls eine Datenquelle für alle Kriterien k mit $w_k > 0$ einen Eintrag $v_{ik} = 0$ hat:

$$\sum_{l=1}^{j} w_l^{'} v_{il} = \sum_{l=1}^{j} 0 = 0$$

Durch die Kombinationen der Gewichte und der Qualitätswerte zwischen 0 und 1 erhalten wir Werte in [0...1]. Wir haben implizit eine Ordnung auf dem j-dimensionalen Raum der Qualitätsvektoren festgelegt.

Beispiel 6. Wir setzen unser Beispiel fort und berechnen die beiden Datenqualitätswerte für S_1 und S_2:

$$DQ(S_1) = 0,75 * 0 + 0,25 * 1 = 0,25$$

$$DQ(S_2) = 0,75 * 1 + 0,25 * 0 = 0,75$$

Wenn wir die SAW-Methode nutzen, um auf Basis der Datenqualität zwischen diesen beiden Quellen zu unterscheiden, wählen wir S_2 aus.

Als nächstes wenden wir uns einer Methode zu, die eine Zielhierarchie verwendet, um einen Wert im Bereich [0...1] zu errechnen. Diese Methode wird allgemein mit "AHP" für "Analytical Hierarchy Process" bezeichnet.

Bei der AHP-Methode gibt der Nutzer keinen Gewichtungsvektor direkt vor. Vielmehr werden alle Ziele in Teilziele zerlegt und der Nutzer gibt auf jeder Ebene der Zielhierarchie in einer Matrix an, wie wichtig ihm diese Teilziele in Relation zueinander sind. Die AHP-Methode berechnet dann Eigenvektoren zu diesen Matrizen und nutzt diese in einer mehrstufigen Gewichtung.

Wir erläutern das Prinzip von AHP zunächst an der in Abbildung 4 dargestellten allgemeinen Zielhierarchie, die wir in Anlehnung an [NM95] vorstellen, bevor wir sie in eine Zielhierarchie für Datenqualität umsetzen. Wir sehen ein Hauptziel, das zunächst aus vier Teilzielen besteht. Davon sind die Teilziele 1

und 3 wiederum in die Unterziele 1.x und 3.x aufgeteilt. Das Unterziel 3.2 ist erneut in drei Teilziele unterteilt. Alle Blätter haben implizit als Nachfolger einen oder mehrere numerische Einträge, die alternative Werte oder Wertkombinationen darstellen[8]. Ist Teilziel 2 Sparsamkeit, so könnte es die zwei möglichen Alternativen Budget 1000 $ oder 100000 $ geben. Die AHP-Methode liefert eine Entscheidung, welche Wertkombinationen insgesamt zu einem guten Ergebnis führen.

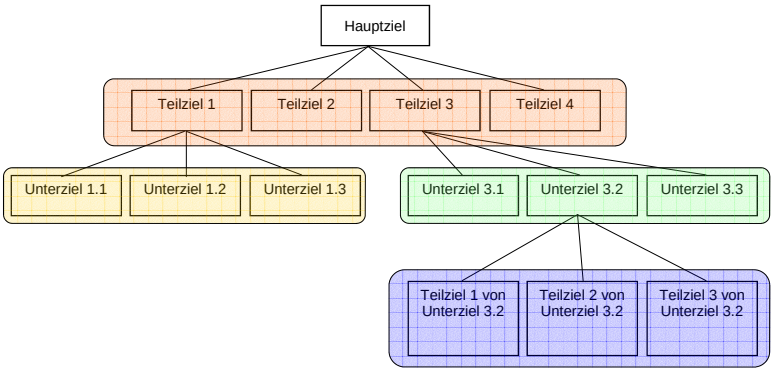

Abbildung 4. Allgemeine AHP-Methode

Dazu berechnen wir auf jeder Ebene für alle Knoten mit gleichem Vaterknoten einen Gewichtungsvektor. In Abbildung 4 haben wir dies durch die graphisch zusammengefassten Knotenmengen (Teilziele 1-4, Unterziele 1.x, Unterziele 3.x, Teilziele x von Unterziel 3.2) angedeutet. Der Nutzer gibt in einer Matrix an, wie wichtig die jeweiligen Ziele in Relation zueinander sind. Dafür vergibt er Werte von 1 (genauso wichtig) bis 9 (sehr viel wichtiger) bzw. von $\frac{1}{9}$ bis $\frac{9}{9}$, um die Umkehrung zu berücksichtigen. Der Nutzer könnte zum Beispiel angeben, dass ihm Teilziel 2 in Relation zu Teilziel 1 etwas wichtiger ist durch Angabe einer 2 in der entsprechenden Matrix.

Im nächsten Schritt muss überprüft werden, ob die Werte, die der Nutzer eingegeben hat, konsistent sind. Inkonsistenzen können leicht entstehen, wie die folgende Matrix $\mathbf{A}$ zeigt:

$$\mathbf{A} = \begin{pmatrix} 1 & 2 & 9 & 2 \\ \frac{1}{2} & 1 & \frac{1}{3} & 2 \\ \vdots & \vdots & \ddots & \vdots \end{pmatrix}$$

[8] Diese Werte könnten als Blätter modelliert werden, hätten aber bei unserer Anwendung mehrere Vorgänger. Daher verzichten wir an dieser Stelle auf diese Modellierung.

Zeile 1 besagt, dass Teilziel 2 etwas wichtiger als Teilziel 1 ist und Teilziel 3 sehr viel wichtiger ist. Damit ergibt sich als abgeleitete Aussage, dass Teilziel 3 wichtiger als Teilziel 2 ist. Andererseits besagt aber die zweite Zeile, dass Teilziel 3 nur $\frac{1}{3}$-mal so wichtig wie Teilziel 2 und somit weniger wichtig ist. Diese beiden Aussagen widersprechen sich. Solche Inkonsistenzen entstehen durch transitive Abhängigkeiten. Bei der AHP-Methode werden sie angezeigt und der Nutzer muss sie korrigieren.

Ist die Matrix konsistent, so hat sie nach [Na02] die Eigenschaft, dass es Vektoren u_i und Zahlen n_i gibt, für die gilt: $\mathbf{A}u_i = n_i u_i$. Das heißt, dass u_i ein Eigenvektor zum Eigenwert n_i ist. Im Allgemeinen gibt es mehrere Eigenvektoren. Für die AHP-Methode wird der Eigenvektor des maximalen Eigenwertes als Gewichtungsvektor genutzt.

Jede ausgehende Kante wird mit dem entsprechenden Eintrag des Eigenvektors markiert. Haben wir in der Situation aus Abbildung 4 für die Teilziele 1 bis 4 den Eigenvektor (t_1, t_2, t_3, t_4) gefunden, so wird an alle Kanten zu den Unterzielen 1.x t_1 geschrieben. Analog werden die Kanten zu Unterzielen 3.x mit t_3 beschriftet. Die Gewichtungsvektoren auf der Ebene der Blätter werden für die Gewichtung der impliziten Werten verwendet, die jedes Blatt hat. Die Werte von Teilziel 2 bzw. Teilziel 4 werden also mit t_2 bzw. t_4 gewichtet.

Schließlich berechnen wir das Gesamtergebnis, indem wir von den Werten unter den Blättern die Pfade bis zur Wurzel hinauf gehen und sie jeweils mit dem Gewichtungswert an den traversierten Kanten multiplizieren. Das Gesamtergebnis ergibt sich aus der Summe der Werte an den eingehenden Kanten beim Hauptziel. So können verschiedene Zusammenstellungen von alternativen Werten zu einem Gesamtergebnis aggregiert und verglichen werden.

Damit können wir, wie in Abbildung 5 gezeigt, eine einfache Transformation auf unser Problem, die Berechnung einen Datenqualitätswertes, vornehmen. Um den Qualitätswert für die Datenquellen 1 und 2 zu berechnen, nehmen wir einmal die nach (2) und (3) skalierten Werte der Datenquelle 1 und berechnen das Gesamtergebnis nach AHP und als Alternative die entsprechenden Werte von Datenquelle 2. Die zwei Qualitätsvektoren der Datenquellen stellen in Abbildung 5 explizit die bisher nur implizit angenommenen alternativen Werte dar. Nach [Na02] führt dieses Vorgehen zu Werten im Bereich [0...1].

Wir bemerken, dass diese Methode für eine einstufige Hierarchie zu einer SAW-Methode wird, bei der lediglich der Gewichtungsvektor über relative Wichtigkeiten und Eigenvektoren bestimmt wird. Abschließend ist zu AHP anzumerken, dass eine höhere Nutzerinteraktion gefordert ist, um eine konsistente Matrix einzugeben, dafür aber die Gewichtung über Paarvergleiche teilautomatisiert erfolgt.

Als letzte Methode stellen wir einen Ansatz vor, der im Kern ein Optimierungsproblem ist. Er führt nicht zu einer Abbildung in den Bereich [0...1], sondern unterscheidet mit mathematischen Methoden zwischen "guten" und "schlechten" Datenquellen. Damit werden Datenquellen nur nach diesen beiden Begriffen unterschieden. Das Verfahren ist als "DEA" für "Data Envelopment Analysis" bekannt.

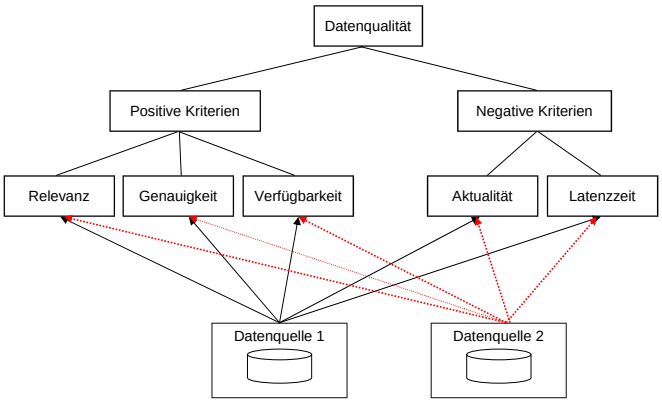

Abbildung 5. AHP-Methode für Berechnung der Datenqualität

Zur Erläuterung des Verfahrens untersuchen wir zunächst einen zweidimensionalen Vektorraum. Dabei können die Einheiten der Dimensionen beliebig sein, so dass wir unsere Qualitätsvektoren nicht skalieren oder normieren müssen. Mit Punkten in einem zweidimensionalen Vektorraum können wir eine maximale Grenze definieren, indem wir diejenigen Punkte verbinden, die am weitesten vom Ursprung entfernt liegen. Damit sind alle Punkte entweder Teile dieser Grenze oder liegen näher am Ursprung als die Grenze. Dieses Vorgehen zeigt Abbildung 6, in der wir zehn Punkte sehen und die dazu gehörige Grenzlinie.

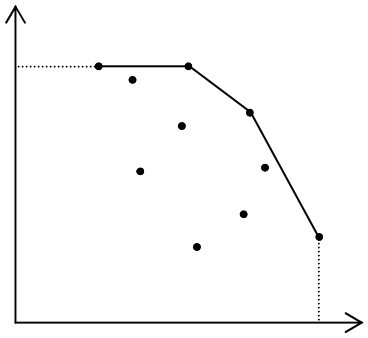

Abbildung 6. Grenzbildung beim DEA-Verfahren

Eine Datenquelle, die auf der Grenzlinie liegt, ist eine gute Datenquelle. Eine Datenquelle die näher am Ursprung liegt als die Grenzlinie ist keine gute, sondern eine schlechte Datenquelle. Eine neu hinzugekommene Datenquelle fällt entweder in eine der beiden Kategorien oder verschiebt die Definition der Grenze. Damit verändert sich unter Umständen die Bewertung von Datenquellen wie Abbildung 7 zeigt.

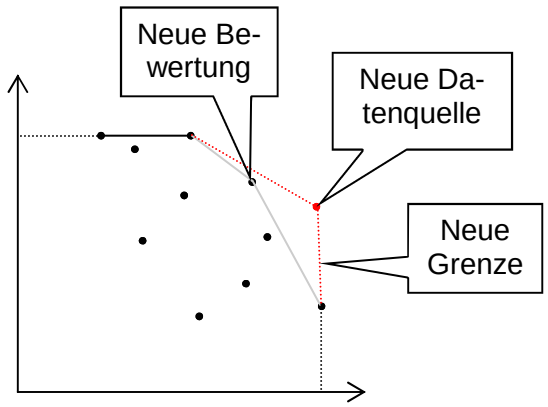

Abbildung 7. Dynamische Anpassung durch neue Datenquellen beim DEA-Verfahren

Nun haben wir bei dieser anschaulichen und intuitiven Vorgehensweise noch nicht erläutert, wie eine konkrete Berechnung aussieht und noch keine unterschiedlichen Gewichtungen der Dimensionen vorgenommen. Dies erfolgt, indem wir das Problem als ein Optimierungsproblem durch ein lineares Programm beschreiben[9]. Wir optimieren dabei für eine Datenquelle S_i die Summe $\sum_k w_k d_{ik}$ mit den Optimierungsparametern $w_k > 0$ und der Nebenbedingung, dass der Datenqualitätswert für alle Quellen ≤ 1 ist.

Die Optimierungsparameter sind die Einträge des Gewichtungsvektors, so dass wir einen optimalen Gewichtungsvektor berechnen. Eingangswerte sind nur die j unskalierten, nicht gewichteten Einträge aller Datenquellen, alle Qualitätsvektoren. Damit erhalten wir den optimalen Wert, den eine Datenquelle erreichen

[9] Gewöhnlich wird erst die mathematische Formulierung als Optimierungsproblem erläutert und daraus die Anschauung mit einer Grenzlinie entwickelt. In dieser Anwendung scheint uns aber das umgekehrte Vorgehen eine intuitive Begriffsbildung zu erlauben.

kann, unter Berücksichtigung der Nebenbedingungen, die die anderen Datenquellen stellen. Diese Nebenbedingungen stellen sicher, dass der Gewichtungsvektor nur so gewählt werden darf, dass alle Datenqualitätswerte im Intervall [0...1] liegen.

Erreicht die Datenquelle den Wert 1, liegt sie auf der oben anschaulich beschriebenen Grenze, erreicht sie einen Wert ≤ 1 liegt sie unterhalb der Grenze. Dieses Optimierungsproblem formulieren wir für die Datenquellen S_i nach [NLF99] und [Na02] als lineares Programm (wir kürzen positives Qualitätskriterium mit p. Q. ab, negatives Qualitätskriterium analog mit n. Q.):

Maximiere Zielfunktion:

$\sum_k$ p. Q. $w_k d_{ik} - \sum_k$ n. Q. $w_k d_{ik}$

Mit Nebenbedingung:

$\sum_k$ p. Q. $w_k d_{lk} - \sum_k$ n. Q. $w_k d_{lk} \leq 1,$ für alle $S_l, w_k \geq \epsilon > 0$

In dem linearen Programm sehen wir, dass wir die negativen Kriterien berücksichtigen, indem wir ihre Anteile an der Zielfunktion negativ betrachten. Außerdem führen wir den Parameter ϵ ein, der es erlaubt, die Bedingung "auf dem Rand" anschaulich in die Bedingung "in der Nähe des Randes" umzuformulieren.

Das DEA-Verfahren setzt Methoden aus der Optimierung zur Ermittlung von Datenqualitätswerten ein. Der Nutzer muss keine Gewichtung vorgeben und als Eingabeparameter werden nur die Qualitätsvektoren der Datenquellen benötigt. Dies ist einerseits für ihn einfacher, andererseits verliert er damit an Einfluss auf den Bewertungs- und damit auf den Auswahlprozess. Der Berechnungsaufwand kann durch Anwendung von Ergebnissen aus der Optimierung begrenzt werden. Durch Anwendung von DEA wird eine Unterscheidung in gute und schlechte Datenquellen vorgenommen, aber es wird kein Ranking ermöglicht. In einem vieldimensionalen Raum mit wenig Datenquellen ist es möglich, dass jede Datenquelle in irgendeiner Dimension die beste ist und dann mit großer Wahrscheinlichkeit zur Grenze gehört. Solche Phänomene müssen bei der Nutzung berücksichtigt werden.

Wir haben drei Methoden vorgestellt, mit denen Qualitätswerte für Datenquellen zu einem Gesamturteil aggregiert werden können. Damit stehen drei Möglichkeiten zur Auswahl, um die Qualität der zur Verfügung stehenden Datenquellen zu beurteilen. Diese Verfahren werden im nächsten Kapitel bei der Bearbeitung von konkreten Anfragen an ein heterogenes Informationssystem benötigt.

6 Qualitätsgetriebene Integration

Wir kommen auf die praktische Problemstellung zurück, die darin besteht, eine Anfrage des Nutzers an das heterogene Informationssystem unter Einbeziehung

der Qualitätsinformationen zu beantworten. Wir zeigen, wie aus dieser Anfrage des Nutzers Anfragen an die Wrapper gewonnen werden, um die benötigten Informationen aus den Datenquellen abzufragen. Diesen Prozess, der einen ausführbaren Anfrageplan erzeugt, nennen wir Query Planning. Wir zeigen zunächst, wie Qualitätsdaten mit einem klassischen Ansatz zum Query Planning verbunden werden können, um die besten Pläne zu ermitteln. Anschließend erläutern wir einen Branch&Bound-Algorithmus, der vollständig auf den Qualitätsdaten beruht, indem diese im Bound-Schritt genutzt werden. Wir beschränken uns darauf, die Grundlagen der Verfahren darzustellen.

6.1 Verbindung von Anfrageplänen und Qualitätsdaten

Die Wrapper modellieren die Datenquellen als relationale Tabellen, die Attribute der Universal Relation beinhalten. Somit können wir die Wrapper als Sichten auf die Universal Relation verstehen. Wir können aus einer Anfrage des Nutzers gegen die Universal Relation alle möglichen Anfragepläne erzeugen, indem wir alle Sichten V_i abfragen, die Attribute enthalten, die in der Anfrage des Nutzers enthalten sind. Dazu verbinden wir alle betroffenen Sichten mit Join-Operatoren über Fremdschlüssel-Beziehungen. Hat der Benutzer Prädikate für Attribute angegeben, übernehmen wir diese bei der Abfrage einer jeden Sicht, die dieses Attribut beinhaltet.

Beispiel 7. Die Universal Relation hat den Primärschlüssel "Firmenname" und die Attribute "Firmenprofil", "Aktienwert" und "Entwicklung". Es sind drei Datenquellen vorhanden. Datenquelle 1 hat zu jeder Firma nur ein Profil gespeichert, Datenquelle 2 nur den aktuellen Kurs und Datenquelle 3 nur die Entwicklung. Alle Datenquellen haben den gleichen Primärschlüssel "Firmenname", der somit als Fremdschlüssel auf die Universal Relation gesehen werden kann.

Der Nutzer möchte Informationen über den Kurs und die Firma. Also müssen wir die Datenquellen 1 und 2 anfragen, die die benötigten Daten, die benötigten Sichten zur Verfügung stellen und die Informationen gemäß dem Primärschlüssel integrieren. Datenquelle 3 fragen wir nicht ab. Der Join erfolgt über das Attribut "Firmenname".

Die Anfragepläne können wir, wie für das vorherige Beispiel in Abbildung 8 gezeigt, als Baumstrukturen modellieren. Da wir davon ausgehen, dass wir die Qualitätsvektoren für die einzelnen Datenquellen kennen, müssen wir angeben, wie wir aus zwei Qualitätsvektoren, die Eingangskanten für einen Join-Operator sind, einen neuen Qualitätsvektor nach dem Join berechnen. Dazu definieren wir zunächst Merge-Funktionen.

Definition 3. *Sei D der Wertebereich für ein Qualitätskriterium. Eine kommutative und assoziative Funktion $\circ : D \times D \rightarrow D$ heißt Merge-Funktion.*

Die Forderungen der Kommutativität und der Assoziativität sind sinnvoll, da der Qualitätswert an der Wurzel nicht von der Reihenfolge der Abfrage abhängen soll, sondern nur von der Qualität der beteiligten Sichten. Müssen wir j Sichten

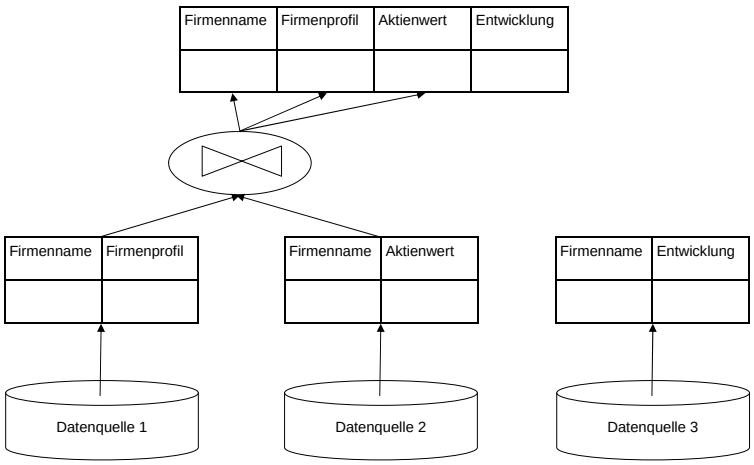

Abbildung 8. Baumstruktur der Anfrage aus Beispiel 7

abfragen, können wir einen Plan A als j-Tupel $(V_1, ..., V_j)$ darstellen. Dies drückt aus, dass wir in irgendeiner Reihenfolge die Sichten V_i über Join-Operatoren verbinden[10].

Für verschiedene Qualitätskriterien sind unterschiedliche Merge-Funktionen erforderlich.

Für die Verfügbarkeit (Wahrscheinlichkeit, dass eine Datenquelle ein Ergebnis liefert) wählen wir das Produkt, da wir dann die Wahrscheinlichkeit erhalten, dass beide Datenquellen verfügbar sind. Für die Antwortzeit nehmen wir das Maximum beider Antwortzeiten, da wir beide Datenquellen parallel abfragen und somit die langsamere Quelle die Gesamtzeit bestimmt.

Sind $Q_L = (d_{L1}, ..., d_{Lb})$ und $Q_R = (d_{R1}, ..., d_{Rb})$ die Qualitätsvektoren des linken und rechten Sohnes eines Join-Knotens, so berechnen wir den Qualitätsvektor (mit Einträgen für b Qualitätskriterien) nach dem Join durch:

$$Q_{L \circ R} = (d_{L1} \circ d_{R1}, ..., d_{Lb} \circ d_{Rb}) \tag{6}$$

Aus den Qualitätsvektoren der Datenquellen leiten wir einen Qualitätsvektor für jede Sicht V_i ab und können durch wiederholte Anwendung von (6) ausgehend von den untersten Elementen im Anfragebaum den Qualitätswert eines Anfrageplans A berechnen.

Wir haben noch nicht berücksichtigt, dass wir davon ausgehen, dass mehrere Datenquellen durch die gleiche Sicht V_i beschrieben werden bzw. dass die gleiche Sicht V_i auf verschiedene Datenquellen zutrifft. Für ein festes i gibt es nicht

[10] Sind die besten Pläne bekannt, können die Joins mit Methoden der Anfrageoptimierung in eine möglichst optimale Reihenfolge gebracht werden. Dies ist aber kein Problem der Datenqualität, für die die Reihenfolge unerheblich ist.

eine einzelne Sicht V_i, sondern eine Menge $\mathcal{V}_i$, die diese Sicht mehrmals aus verschiedenen Datenquellen enthält[11].

Als Konsequenz gibt es nicht mehr einen Anfrageplan A, sondern eine Menge $\mathcal{A}$ von Anfrageplänen. Müssen wir j Sichten $V_1, ...V_j$ abfragen, so sind alle Anfragepläne *korrekt*, die aus jeder der Mengen $\mathcal{V}_i$ ein Element, eine Sicht V_i, auswählen und verbinden[12]. Wir wollen aber nicht alle $|\mathcal{A}|$ korrekten, sondern in drei Schritten die qualitativ besten N Anfragepläne ausführen.

Im ersten Schritt reduzieren wir die Komplexität, indem wir schlechte Datenquellen aussortieren und nicht mehr berücksichtigen. Wir berechnen mit der DEA-Methode alle schlechten Datenquellen ($DEA(S_i) < 1$) unter Berücksichtigung der Qualitätskriterien, die nur von der Datenquelle abhängen. Als Ausnahme lassen wir eine schlechte Datenquelle S_j zu, falls sie ein Attribut bereitstellt, dass in keiner anderen Datenquelle $S_{k,k\neq j}$ gespeichert ist. Dann entfernen wir alle Sichten von schlechten Datenquellen. Wir versuchen somit $|\mathcal{V}_i|$ für alle i zu reduzieren. Diesen Schritt müssen wir nur einmal durchführen, da die Auswahl solange gültig ist, bis sich die Qualität einer Datenquelle stark ändert.

Im zweiten Schritt berechnen wir die Menge $\mathcal{A}$ aller korrekten Anfragepläne. Damit enthält die Menge $\mathcal{A}$ alle möglichen Kombinationen von Sichten $(V_1, ..., V_j), V_i \in \mathcal{V}_i$.

Im dritten Schritt berechnen wir für alle Pläne gemäß (6) den Qualitätswert. Anschließend nutzen wir SAW, um den Plänen einen Datenqualitätswert zuzuordnen und sie entsprechend zu sortieren. Dann führen wir die N besten Pläne aus.

Auf diese Weise können Qualitätsinformationen mit Anfragepläne verbunden werden, unabhängig davon, wie die Menge $\mathcal{A}$ der Anfragepläne berechnet wird.

6.2 Integration von Qualitätsdaten in die Anfragebearbeitung

Als Alternative zu dem in 6.1 dargestellten Verfahren skizzieren wir eine Möglichkeit, wie nach [Na02] Qualitätsdaten in einem Branch&Bound-Algorithmus Kern des Query Plannings werden. Wir führen zunächst den ersten Schritt wie in 6.1 vorgestellt aus, sortieren also schlechte Datenquellen mit einer DEA-Analyse aus.

In unserem Modell gehen wir wiederum davon aus, dass wir Sichten auf die Universal Relation betrachten und j Mengen von Sichten, $\mathcal{V}_1, ...\mathcal{V}_j$ abfragen müssen. Jeder vollständige Ausführungsplan ist ein Tupel $(V_1, ..., V_j)$, das alle j Mengen berücksichtigt. Wir wenden einen Standard-Branch&Bound-Algorithmus an.

Haben wir im k-ten Schritt die Mengen $\mathcal{V}_1, ..., \mathcal{V}_k$ berücksichtigt, so besteht der Branch-Schritt darin, dass wir erweiterte Pläne erzeugen, indem wir Sichten

[11] Stellen z Datenquellen die Sicht V_i bereit, müssten wir in einer genauen mathematischen Beschreibung $V_{i1}, ..., V_{iz} \in \mathcal{V}_i$ schreiben. Der Übersicht wegen verzichten wir auf die doppelten Indizes und schreiben V_i für ein beliebiges Element aus $\mathcal{V}_i$.

[12] Hier müssten wir in der genauen mathematischen Notation einen Anfrageplan als j-Tupel formulieren, das aus beliebigen Elementen der $\mathcal{V}_i$ besteht: $(V_{1l}, ...V_{jm})$.

$V_{k+1} \in \mathcal{V}_{k+1}$ neu berücksichtigen. Sei $P = (V_1, ..., V_k, V_{k+1})$ ein solcher erweiterter Plan.

Nun müssen wir im Bound-Schritt für diesen erweiterten Plan P eine obere Schranke finden, die angibt, welchen Qualitätswert er maximal erreichen wird, egal, wie er zu einem vollständigen Plan $P^{'}$ fortgesetzt wird:

$$P^{'} = (V_1, ..., V_k, V_{k+1}, V_{k+2}, .., V_j)$$

Wir kennen den bisherigen Qualitätsvektor im k-ten Schritt, denn dieser berechnet sich aus Kenntnis der Qualitätsvektoren der ersten k Sichten nach (6) zu:

$$IQ(P) = IQ(V_1) \circ ... \circ IQ(V_k) \circ IQ(V_{k+1})$$

Wir müssen für die Qualitätsvektoren $IQ(V_{k+2}), ..., IQ(V_j)$ dieser noch nicht ausgewählten Sichten eine obere Schranke finden.

Betrachten wir eine Menge von Sichten $\mathcal{V}_i$, so konstruieren wir einen beschränkenden Qualitätsvektor, indem wir für ein Qualitätskriterium a den besten Wert annehmen, den irgendeine Sicht in $V_i \in \mathcal{V}_i$ für a hat. Diesen maximalen Wert bezeichnen wir mit d_{ia}^{max}. Dann ist eine obere Schranke $\top$ gegeben durch:

$$\top(\mathcal{V}_i) := (d_{i1}^{max}, ..., d_{ib}^{max})$$

Die Eigenschaft, dass $\top$ eine obere Schranke für $\mathcal{V}_i$ ist, ist für $\top(\mathcal{V}_i)$ erfüllt, da keine Sicht $V_i \in \mathcal{V}_i$ irgendeinen dieser Einträge übertreffen kann.

Für eine beliebige Fortsetzung des erweiterten Plans P zu einem vollständigen Plan $P^{'} = (V_1, ..., V_k, V_{k+1}, V_{k+2}, .., V_j)$ nutzen wir $\top(\mathcal{V}_{k+2}), ..., \top(\mathcal{V}_j)$ als obere Schranke $\top(P^{'})$:

$$\top(P^{'}) = IQ(V_1) \circ ... \circ IQ(V_k) \circ IQ(V_{k+1}) \circ \top(\mathcal{V}_{k+2}) \circ ... \circ \top(\mathcal{V}_j)$$

Die Eigenschaft, dass dies tatsächlich für beliebige Planfortsetzungen als obere Schranke gültig ist, beweist Naumann in [Na02][13].

Dieser Vorgehen liefert uns den besten Plan und wir erweitern es einfach auf die besten N Pläne, indem wir im Bound-Schritt den Wert des N-ten, bereits gefundenen, vollständigen Plans betrachten. Ist dieser Wert höher als die Schranke eines unvollständigen Plans P, brauchen wir P nicht mehr weiter zu berücksichtigen, da er niemals zu den N besten Plänen gehören kann.

Da wir einen Branch&Bound-Algorithmus nutzen, können Resultate aus der Algorithmik wie die effiziente Implementierung von Warteschlangen, die nach Prioritäten sortiert sind, direkt verwendet werden.

7 Zusammenfassung

Die Datenqualität hat eine wichtige Stellung in heterogenen Informationssystemen. Ihre Beurteilung ist der Schlüssel für hochwertige Ergebnisse auf Nutzeranfragen. Wir haben in dieser Ausarbeitung zunächst den Begriff der Datenqualität genauer gefasst und wichtige Qualitätskriterien vorgestellt, die alle Teil der

[13] Theorem 5.3.1 in [Na02]

Datenqualität sind. Wir haben die Qualitätskriterien unter den vier Bereichen inhaltliche, technische, intellektuelle und präsentationsbezogene Qualitätskriterien systematisiert. Für konkrete Anwendungen haben wir darauf verwiesen, dass, abhängig von der Anwendungsdomäne, eine Auswahl aus den vielen Kriterien getroffen werden muss.

Nach der Identifikation dieser Qualitätskriterien stand die Datenerhebung im Vordergrund. Dazu haben wir im Nutzer, der Datenquelle an sich und dem Prozess der Anfrageverarbeitung wichtige Quellen für die Qualitätsdaten gefunden. Damit steht eine weitere Systematisierung nach diesen drei Kategorien zur Verfügung.

Im Anschluss haben wir eine mathematische Modellierung der Datenqualität bezogen auf j Kriterien in einem j-dimensionalen Vektorraum vorgenommen. Dabei haben wir zwei Methoden vorgestellt, die Qualitätsvektoren zu normalisieren. Ausgehend von diesen Grundlagen haben wir mit SAW, AHP und DEA drei Methoden vorgestellt, Datenquellen ausgehend von ihren Qualitätsvektoren zu beschreiben.

Zum Abschluss der Ausarbeitung sind wir darauf eingegangen, wie wir die Beschreibung von Qualitätsvektoren beim Query Planning nutzen können. Dabei haben wir gezeigt, dass es problemlos möglich ist, einen beliebigen vorhandenen Algorithmus mit Qualitätsdaten zu ergänzen. Mit der Vorstellung eines Branch&Bound-Algorithmus, der die Datenqualität im Bound-Schritt nutzt, steht eine elegante Methode zur Verfügung, die Datenqualität zur Grundlage des Query Planning und damit der Anfragebearbeitung im Mediator zu machen.

Bei allen Kapiteln haben wir eine Auswahl aus den in der angegebenen Literatur behandelten Themen vorgenommen. Zu allen Problemen stehen noch weitere Methoden zur Verfügung. Den Kern der Ausarbeitung bildet die Systematisierung des Begriffes der Datenqualität. Diese Systematisierung wird durch die Ersetzung von Methoden nicht in Frage gestellt. Sie kann einfach auf eine Betrachtung von Systemen außerhalb des genutzten Architektur-Modells, wie in Peer-to-peer-Netzwerken oder in Global-as-view-Ansätzen, übertragen werden. Weitere Aspekte ergeben sich, wenn man davon ausgeht, dass Datenquellen versuchen, die Qualitätsdaten zu manipulieren. Dies kann mit dem Ziel geschehen, den eigenen Wert zu steigern oder zu erreichen, dass andere Quellen zu schlecht bewertet werden. Für praktische Anwendungen müssen solche Manipulationen erkannt werden.

Literatur

[Zi94] Zink, Klaus J. (Hrsg.): Qualität als Managementaufgabe: Total Quality Management. Verlag Moderne Industrie, Landsberg, 3. Auflage 1994.

[Ma99] Masing, Walter (Hrsg.): Handbuch Qualitätsmanagement. Carl Hanser Verlag München Wien, 4. Auflage 1999.

[SSC03] De Santis, Luca, Scannapieco, Monica, Catarci1, Tiziana: Trusting Data Quality in Cooperative Information Systems. In: LNCS 2888, 2003, pp. 354-369.

[Os01] Ostländer, Nicole: Daten vs. Information: Eine vergleichende Qualitätsanalyse unterschiedlicher Datengrundlagen zur Bestimmung des Stoffaustrags ausgewählter Einzugsgebiete im Bergischen Land. Diplomarbeit an der Westfälischen Wilhelmsuniversität Münster, Fachbereich Geowissenschaften.

[We06] Webbaromter auf www.webhits.de. Zuletzt besucht 14.05.2006.

[Na02] Naumann, Felix: Quality-Driven Query Answering for Integrated Information Systems. Dissertation der Humboldt Universität zu Berlin. In: LNCS 2261, 2002.

[NR00] Naumann, Felix, Rolker, Claudia: Assessment Methods for Information Quality Criteria. In: Proceedings of the MIT Conference on Information Quality, 2000, pp. 148-162.

[De03] Dessloch, S.: Vorlesung Grundlagen Betrieblicher Informationssystem, Technische Universität Kaiserslautern, Kapitel 2, Sommersemester 2003.

[NM95] Norris, Gregory A., Marshall, Herold E., Multiattribute Decision Analysis Method for Evaluating Buildings and Building Systems, U.S. Department of Commerce, Office of Applied Economics, Building and Fire Research Laboratory, National Institue of Standards and Technology, Gaithersburg 1995.

[NLF99] Naumann, Felix, Leser, Ulf, Freytag, Johann Christoph, Quality-driven Integration of Heterogeneous Information Systems, 1999.