Ein Konzept zur automatisierten Klassifizierung von Informationen für das Information Lifecycle Management – dargestellt am Beispiel des SAP NetWeaver Business Intelligence

Monique Kosler*, Mauricio Matthesius**, Dirk Stelzer***

*IBM Deutschland
Rathausstraße 7
09111 Chemnitz
Monique.Kosler@de.ibm.com

**sones GmbH
Eugen-Richter-Straße 44
99085 Erfurt
mauricio@sones.de

***Fachgebiet Informations- und Wissensmanagement
TU Ilmenau
Postfach 100565
98684 Ilmenau
Dirk.Stelzer@tu-ilmenau.de

Abstract: Information Lifecycle Management (ILM) hat zum Ziel, Informationen zu klassifizieren und entsprechend ihrer Klasse auf dem jeweils günstigen Speichermedium zur Verfügung zu stellen. Unser Beitrag konzentriert sich auf die zentrale Teilaufgabe des ILM, die Informationsklassifizierung. In der Unternehmenspraxis muss diese zumindest teilweise automatisiert werden. Zunächst ordnen wir die Informationsklassifizierung mit Hilfe eines Vorgehensmodells in das ILM ein. Auf der Grundlage der Vorarbeiten von Chen entwickeln wir ein Konzept zur automatisierten Klassifizierung von Informationen in einem Data Warehouse. Anhand einer prototypischen Implementierung im Rahmen eines SAP-BI-Projektes demonstrieren wir die praktische Anwendung des Konzeptes und erörtern, welche Verbesserungsmöglichkeiten dabei identifiziert wurden.

1 Einleitung

Ein großer Teil der strukturierten Informationen eines Unternehmens befindet sich innerhalb von Datenbanken, beispielsweise in Data Warehouses [Sh06]. Im Data-Warehouse-Bereich ist mit einem starken Informationswachstum zu rechnen, da die Bedeutung von Data Warehouses für Unternehmen stark zunimmt [Sha06], [WW07].

Das Informationswachstum führt zu hohen Speicher- und Administrationskosten [Ab06], [Al01], [TS07]. Die Kosten für die Administration der Datenträger bzw. Datenspeicher, wie beispielsweise die Datensicherung und Spiegelung der Informationen sowie die sichere Aufbewahrung, sind um das vier- bis achtfache höher als die Beschaffungskosten der Datenträger [Gr03] selbst. Zudem müssen Unternehmensinformationen aufgrund der gesetzlichen Anforderungen für längere Zeiträume sicher aufbewahrt werden [Sh06], was zu einer zusätzlichen Erhöhung der Kosten führt.

Information Lifecycle Management (ILM) zielt darauf ab, Informationen[1] zu bewerten, zu klassifizieren und kostengünstig zu verwalten [MHP05], [TS07]. Informationen, die einen höheren Wert für ein Unternehmen aufweisen, müssen von Informationen mit geringerem Wert getrennt [Sh06], [MW99] und den Unternehmensanforderungen entsprechend auf dem jeweils sinnvollsten Speichermedium bereitgestellt und verwaltet werden [Ab05], [Bh05], [MHP05], [Mo06]. Als schwierig erweist sich die Bewertung und anschließende Klassifizierung von Informationen [Ab06], [Bh05], [Ch05], [Me04], [TS07], [YB02]. Hierfür werden zahlreiche Angaben benötigt, wie beispielsweise die Anzahl der Lese- und Schreibzugriffe auf diese Objekte [Ch05], [El03]. Zusätzlich sind gesetzliche Anforderungen zur Aufbewahrung der Informationen [Sh06] sowie Gewohnheiten der Nutzer [Me04], [Sha06] hinsichtlich der Informationsnutzung von Relevanz. Die Erhebung, kontinuierliche Aktualisierung und Auswertung dieser Bewertungsinformationen ohne automatisierte Unterstützung ist praktisch kaum möglich und wirtschaftlich nicht sinnvoll. Aus diesem Grund muss die Informationsklassifizierung zumindest teilweise automatisiert erfolgen.

In diesem Beitrag entwickeln wir ein Konzept zur automatisierten Klassifizierung von Informationen für das Information Lifecycle Management und stellen dessen Anwendung beispielhaft an SAP NetWeaver Business Intelligence (SAP BI) dar. Für die Bewertung bzw. Klassifizierung von Informationen in Data-Warehouse-Systemen sind unseres Wissens bisher keine vergleichbaren Konzepte publiziert worden.

2 Vorgehensmodell des ILM

Anhand des in Abbildung 1 dargestellten Vorgehensmodells [MS08] werden zunächst wesentliche Funktionen des ILM kurz beschrieben.

[1] Tatsächlich werden im Rahmen des ILM Daten bzw. Dateien analysiert und bewertet. Da in den meisten Publikationen der Betrachtungsgegenstand jedoch als Information bezeichnet wird, verwenden auch wir diesen Begriff.

Abbildung 1: Vorgehensmodell des ILM

In Phase eins, „Analyse der Systemlandschaft“, werden in der Teilphase „Ermittlung der Systeme“ zunächst die in einem Unternehmen vorhandenen Data-Warehouse-Systeme ermittelt und untersucht. Dabei sind insbesondere die Verbindungen[2] zwischen den Systemen sowie die durch die Systeme verarbeiteten Informationen von Interesse, die in den Teilphasen „Ermittlung der Verbindungen“ und „Ermittlung der verarbeiteten Informationen“ identifiziert werden.

Die physische Größe, die Anwender und das Informationswachstum der Data-Warehouse-Systeme werden in Phase zwei, „Untersuchung relevanter Systeme“, ermittelt. Größe und Wachstum sind wichtige Indikatoren dafür, welche Systeme im Hinblick auf eine Kostenreduktion zuerst untersucht werden sollten. Des Weiteren werden die Anwender der jeweiligen Systeme ermittelt, da diese für die dritte Phase, die „Informationsklassifizierung“, von Bedeutung sind. In der Teilphase „Festlegung von Klassifikationskriterien“ werden für die Informationsklassifizierung relevante Kriterien festgelegt, wie beispielsweise die Zugriffshäufigkeit auf die Informationen sowie das Anwender- und Administratorenwissen über die Systeme und die darin enthaltenen Informationen. Vorgehensweise sowie Methoden zur Messung der relevanten Kriterien werden in der Teilphase „Operationalisierung von Messmethoden und -indikatoren“ bestimmt. Anschließend erfolgt die „Erhebung der Messwerte“ und die „Interpretation und Dokumentation der Ergebnisse“. In diesen beiden Teilphasen werden die Informationen entsprechend der festgelegten Kriterien und Methoden bewertet und klassifiziert.

[2] Hierbei ist von besonderem Interesse, welche Data-Warehouse-Systeme Informationen aus anderen Data-Warehouse-Systemen aufnehmen, verarbeiten und erneut an andere Data-Warehouse-Systeme weitergeben. So wird deutlich, ob bestimmte Informationen auf unterschiedlichen Systemen redundant gehalten werden und ob eine mehrfache Archivierung dieser Informationen innerhalb der folgenden Phasen des Vorgehensmodells notwendig und sinnvoll ist.

In der Literatur werden unterschiedliche Kriterien zur Bewertung von Informationen herangezogen. Nach unserer Auffassung eignen sich beispielsweise die erzielten monetären Erträge, die aus der Nutzung der Informationen entstehen, oder die entstandenen Kosten, die bei der Sammlung, Akquisition und Erstellung der Informationen anfallen [MW99], eher schlecht. Aufgrund der in der Regel komplexen Verwendung von Informationen lässt sich ein direkter Bezug zu Umsatz- oder Kostengrößen nur sehr schwer herstellen. Wir bevorzugen den Nutzungsgrad als Kriterium für die automatisierte Bewertung von Informationen im Rahmen des ILM. Der Nutzungsgrad gibt an, in welchem Maße eine Information zur Unterstützung von Entscheidungen herangezogen wurde. Er bestimmt sich danach, wie oft und wann der Zugriff auf eine Information erfolgte. Wie bei den alternativ genannten Kriterien kann der Nutzungsgrad Hinweise zur Bedeutung einer Information liefern. Im Gegensatz zu den erstgenannten Kriterien lässt er sich aber ohne großen Aufwand ermitteln und automatisiert messen [Bh05], [Ch05], [CGY07], [Sha06]. Der Wert einer Information resultiert im Rahmen unseres Beitrags daher aus der Quantifizierung des Nutzungsgrades[3]. Dadurch wird eine berechenbare Grundlage zur anschließenden Klassifizierung der Informationen geliefert [Ch05], [Do04], [Sha06].

Eine Klasse bezeichnet das geeignete Speichermedium [MHP05], auf welches die Informationen verlagert werden sollen. Hierbei werden die Informationen in vier Klassen eingeteilt:

- Klasse „Online“: Informationen dieser Klasse werden oft von Anwendern für Analysen benötigt und sollten daher für den sofortigen und schnellen Zugriff [Pe05] im Online-Speicherbereich vorgehalten werden.
- Klasse „Nearline“: Informationen, auf die in regelmäßigen Abständen zugegriffen wird, für die aber eine Ablage auf den kostenintensiven Speichermedien des Data-Warehouse-Systems nicht lohnend ist, werden auf Nearline-Storage-Systemen [LLZ04], [MHP05] abgelegt.
- Klasse „Offline“: Informationen, auf die nur in Ausnahmefällen zugegriffen wird oder deren Aufbewahrung aufgrund rechtlicher Anforderungen notwendig ist, können auf kostengünstigen Speichermedien langfristig aufbewahrt werden [KSH96].[4]
- Klasse „Löschen“: Auf Informationen dieser Gruppe wird nicht mehr zugegriffen. Sie haben keinen Wert für das Unternehmen und können gelöscht werden, sofern keine rechtlichen Aufbewahrungsrestriktionen bestehen.

In Phase vier, „Informationsverlagerung“, werden die Informationen entsprechend ihrer Klasse auf die dafür vorgesehenen Speichermedien verlagert.

---

[3] Vgl. Abschnitt 4.

[4] Hierfür können Magnetbänder verwendet werden [Ja98]. Die Speicherung auf Magnetbändern gilt als kostengünstigste Alternative [HS96].

3 Anforderungen an Konzepte zur automatisierten Klassifizierung von Informationen

Mit Hilfe einer Literaturanalyse haben wir verschiedene Anforderungen an Konzepte zur automatisierten Informationsklassifizierung identifiziert. Diese werden im Folgenden vorgestellt.

Zunächst sollten Konzepte zur automatisierten Klassifizierung von Informationen die Kriterien Zugriffshäufigkeit und Zugriffszeitpunkt zur Bestimmung des Nutzungsgrades der Informationen verwenden. Ein Zugriff erfolgt, wenn eine Information gelesen oder verändert wird [Sha06]. Etwa sechzig bis achtzig Prozent der Informationen eines Unternehmens bleiben ungenutzt. Auf sie wird nur noch selten zugegriffen [Ch05], [Sha06], [Sh06], [Za04], was eine Auslagerung auf kostengünstigere Speicherbereiche rechtfertigt. Weiterhin muss ermittelt werden, zu welchen Zeitpunkten auf eine Information zugegriffen wurde, um zu bestimmen, ob Informationen vor kürzerer Zeit genutzt wurden oder ob für einen längeren Zeitraum nicht darauf zugegriffen worden ist [Ch05].[5]

Für die Automatisierung müssen darüber hinaus Funktionen implementiert werden, welche die Informationen anhand des Nutzungsgrades bewerten und anschließend den zugehörigen Klassen zuordnen [Ch05], [Me04], [Sha06], [Ve05]. Die Bewertungsfunktionen müssen in der Lage sein, basierend auf der Historie der Zugriffshäufigkeiten und -zeitpunkte, die zukünftigen Zugriffshäufigkeiten und -zeitpunkte zu prognostizieren bzw. zu berechnen [Ab06], [Do04], [Me04], [Sha06]. Damit wird die automatisierte Informationsklassifizierung vereinfacht, da Zugriffshäufigkeiten und -zeitpunkte sowie Klassenzugehörigkeiten nicht ständig neu berechnet werden müssen, sondern eine Vorausplanung bezüglich eines definierten Zeitraums erfolgen kann. Dies wirkt sich positiv auf die Performanz der Data-Warehouse-Systeme und damit auf die Kosteneinsparungspotentiale aus.

Trotz der Automatisierung sollte das Wissen der Anwender und Administratoren über die Systeme und Informationen berücksichtigt werden [Bh05], [Do04], [Pa05]. Anwender und Administratoren können wichtige Hinweise zur Verwendung der Informationen geben, beispielsweise auf Basis der aktuellen und zukünftigen Projekt- bzw. Auftragslage oder der Verwendung von Informationen innerhalb verschiedener Abteilungen des Unternehmens. Insbesondere bei der Definition der Bewertungsfunktionen und bei der Festlegung, ab welcher Zugriffshäufigkeit eine Information einer bestimmten Klasse angehören soll, ist dies von Relevanz. Außerdem ist das Wissen der Anwender und Administratoren für die Validierung der Klassifikation wichtig. Konzepte zur automatisierten Informationsklassifizierung sollten deshalb die Möglichkeit bieten, die Klassenzugehörigkeit auch nach der automatischen Klassifizierung manuell anzupassen. Dies würde es ermöglichen, nicht sinnvoll erscheinende Klassenzuordnungen noch vor der Informationsverlagerung zu revidieren. Dies ist insbesondere bei Vorschlägen zur Löschung von Informationen wichtig.

---

[5] Vgl. Abschnitt 4.3. Dazu müssen durch das Data-Warehouse-System entsprechende Statistiken bzw. Metadaten protokolliert werden.

Neben dem Anwender- und Administratorenwissen sollte auch die Berücksichtigung rechtlicher Rahmenbedingungen ein Bestandteil der nachträglichen Validierung der Klassenzugehörigkeit sein, womit eine weitere wichtige Anforderung an Konzepte zur automatisierten Klassifizierung von Informationen identifiziert werden kann [MB07], [Sh06]. Muss eine Information aufgrund rechtlicher Restriktionen aufbewahrt werden, so darf sie nicht gelöscht werden, obwohl eventuell für einen längeren Zeitraum nicht auf sie zugegriffen wurde. Andererseits existieren Informationen, die nach einem gesetzlich vorgeschriebenen Zeitraum gelöscht werden müssen, obwohl oft darauf zugegriffen worden ist. In beiden Fällen ist eine nachträgliche Anpassung der Klassenzugehörigkeit der betreffenden Information nach einer automatisierten und auf Grundlage des Nutzungsgrades vorgenommenen Klassifikation notwendig.

Wesentliches Ziel von ILM ist die Einsparung von Kosten. Die automatisierte Informationsklassifizierung sollte daher nicht zuletzt zu einer Reduktion der Administrationskosten für Datenspeicher [Me04], [Sha06] führen. Weiterhin muss eine Reduktion der Aufbewahrungskosten aufgrund der Speicherung der Informationen auf dem jeweils optimalen Speichermedium erfolgen [Sha06], damit von einem effektiven ILM gesprochen werden kann.

Es sind bereits verschiedene Konzepte zur automatisierten Klassifizierung von Informationen publiziert worden. Verma et al. [Ve05] und Mesnier et al. [Me04] verwenden als Klassifikationskriterium den Typ der Information bzw. den Dateityp. Informationen gleichen Dateityps haben ihrer Meinung nach dieselben Eigenschaften und können einer gemeinsamen Klasse zugeordnet werden [Me04], [Ve05]. Zadok et al. [Za04] sowie Chandra, Gehani und Yu [CGY07] konzentrieren sich stark auf das Anwender- und Administratorenwissen über Data-Warehouse-Systeme und Informationen. Jeder Anwender und Administrator erzeugt eigene Bewertungsfunktionen bzw. bewertet und klassifiziert seine Informationen selbst. Shah et al. [Sha06] und Bhagwan et al. [Bh05] verwenden die Zugriffshäufigkeit und berücksichtigen zusätzlich das Anwender- und Administratorenwissen sowie die Performanz der Data-Warehouse-Systeme.

Chen [CH05] bezieht sich auf den Nutzungsgrad von Informationen und zieht zur Bewertung die Kriterien Zugriffshäufigkeit und Zugriffszeitpunkt heran. Bezüglich der Entwicklung unseres Klassifizierungskonzeptes orientieren wir uns an den Vorschlägen von Chen, welche im Folgenden vorgestellt werden. Im weiteren Verlauf des Beitrages werden diese Vorschläge auf den Data-Warehouse-Bereich übertragen und am Beispiel des SAP BI beispielhaft umgesetzt. Abschließend wird untersucht, inwieweit der von uns entwickelte Ansatz zur automatisierten Informationsklassifizierung in Data-Warehouse-Systemen die in diesem Abschnitt vorgestellten Anforderungen erfüllt.

4 Konzept zur automatisierten Klassifizierung von Informationen eines Data Warehouse

4.1 Bewertung von Informationen anhand des Verfahrens nach Chen

Das Verfahren nach Chen bestimmt einen Wert für Informationen zu einem bestimmten Zeitpunkt.[6] Der ermittelte Wert basiert auf den Statistiken einer zuvor festgelegten Betrachtungsperiode. Diese Statistiken zeichnen auf, wann und wie oft ein Zugriff auf die betrachteten Informationen erfolgte. Daraus lassen sich Rückschlüsse über den Nutzungsgrad der Informationen ziehen.

Für die Bestimmung eines Wertes für Informationen auf Grundlage der Kombination von Zugriffshäufigkeit und Zugriffszeitpunkten teilt Chen die Betrachtungsperiode in Phasen gleicher Länge ein. Die Aufzeichnung der Statistiken erfolgt separat für jede Phase. Die Zugriffe auf Informationen innerhalb einer Phase sind gleichgewichtet. Für jede Phase wird ein Gewicht bestimmt. Je höher das Gewicht einer Phase ist, desto mehr trägt deren Zugriffsstatistik zur Bestimmung des Wertes einer Information bei. Hierdurch wird die Berücksichtigung der Zugriffszeitpunkte bei der Wertbestimmung erreicht. Die Summe der Gewichte aller Phasen der Betrachtungsperiode ist 1.

Die Länge der Phasen bestimmt, wie stark das Verfahren sich an den Zugriffszeitpunkten oder der Zugriffshäufigkeit ausrichtet. Ausgedehnte Phasen reduzieren den Effekt der Berücksichtigung von Zugriffszeitpunkten. Die Wertbestimmung erfolgt dann vorwiegend aufgrund der Zugriffshäufigkeit. Mit Hilfe einer Fallstudie hat Chen ermittelt, dass die Länge der Phasen bei einer Betrachtungsperiode von 60 Tagen zwischen 8 und 16 Tagen liegen sollte [Ch05]. In dieser Spanne ist das Verfahren robust gegenüber der Länge der Phasen.

Ein Zähler i verleiht jeder Phase einen Zahlenwert. Die Zählung beginnt in der Gegenwart. Weniger weit zurückliegende Phasen haben kleine Zahlenwerte, weit zurückliegende Phasen weisen höhere Zahlenwerte auf. Die Betrachtungsperiode (vp-valuation period) beginnt vor und endet mit dem Zeitpunkt t, zu dem der Wert der Informationen bestimmt werden soll. Die Anzahl der Phasen ist n, deren Länge in Tagen wird durch s bezeichnet.

$$vp = \left[t - (n \cdot s),\ t\right]$$

Unter der Bedingung, dass Informationen, auf die erst vor kurzem zugegriffen wurde, einen höheren Wert haben als Informationen, die vor längerer Zeit letztmalig genutzt worden sind, sollten die Gewichte der Phasen von der Vergangenheit zur Gegenwart hin steigen. Chen schlägt dazu folgende Berechnungsformel vor:

---

[6] Vgl. Abschnitt 2, Vorgehensmodell Phase 3.

$$w_i = \frac{\left(\frac{1}{x}\right)^i}{\sum_{j=1}^{n}\left(\frac{1}{x}\right)^j} \quad mit \quad x \geq 1$$

Hierbei stellt w_i das Gewicht der i-ten Phase dar. Der Parameter x ist frei wählbar. Dieser beeinflusst die Gewichtsverteilung der Phasen in der Betrachtungsperiode. Je größer x gewählt wird, desto steiler ist die Gewichtsverteilung. Chen betont, dass sehr flache oder sehr steile Gewichtsverteilungen vermieden werden sollten. Eine flache Gewichtsverteilung bei einem niedrigen x-Wert führt im Extremfall zur Gleichwertigkeit aller Phasen. Das Verfahren richtet sich dann vornehmlich an der Zugriffshäufigkeit aus. Eine besonders steile Gewichtsverteilung bei einem hohen Wert für x führt zur Vernachlässigung der Zugriffshäufigkeiten, da tendenziell alle Informationen, auf die in letzter Zeit zugegriffen worden ist, einen hohen Wert erhalten. Mit einem Wert für x zwischen zwei und drei ist das Verfahren laut Chen in der von ihr untersuchten Fallstudie hinreichend robust.

Der Wert einer Information ergibt sich aus der Multiplikation des Phasen-Gewichtes mit der Anzahl der Zugriffe innerhalb einer Phase, aufsummiert über alle Phasen der Betrachtungsperiode. Zur Anwendung dieses Bewertungsansatzes im Data-Warehouse-Bereich, schlagen wir folgendes Vorgehen vor.

4.2 Übertragung des Verfahrens nach Chen auf den Data-Warehouse-Bereich

Den Ausgangspunkt für die Klassifizierung von Informationen[7] in einem Data Warehouse bilden multidimensionale Datenwürfel. Auf einem Datenwürfel sind Abfragen definiert, die von den Nutzern des Data-Warehouse-Systems ausgeführt und deren Ergebnisse für die Unterstützung von Entscheidungen herangezogen werden. Für jede Abfrage wird entsprechend des Ansatzes nach Chen ein Wert bestimmt. Dieser Wert wird auf alle Informationen (Datensätze) der Faktentabelle des Datenwürfels übertragen, die von der Abfrage angesprochen werden. Dafür wird unserem Klassifizierungskonzept eine relationale Datenbankarchitektur zugrunde gelegt. Im Ergebnis sind alle Informationen der Faktentabelle mit einem Wert hinterlegt, aufgrund dessen Rückschlüsse über den Nutzungsgrad der einzelnen Informationen gezogen und eine geeignete Klassifizierung vorgenommen werden können.

[7] Im Kontext des Data Warehouse sind unter Informationen die in den multidimensionalen Datenwürfeln befindlichen Datensätze zu verstehen.

Die Beschränkung auf die Bewertung der Informationen aus der Faktentabelle eines Datenwürfels ist auf die Tatsache zurückzuführen, dass sich in den Faktentabellen der größte Teil des Informationsvolumens eines Data Warehouse befindet. Die Auslagerung von Informationen der Dimensionstabellen würde nicht nur zu Komplikationen bei der Durchführung von Auswertungen führen, sondern zudem keinen nennenswerten Beitrag zur Beschränkung des Informationsvolumens im Online-Speicherbereich im Rahmen eines effektiven ILM leisten.

Um den Wert für eine Abfrage d zum Zeitpunkt t ($v_t(d)$) zu bestimmen, wird nun analog zum in Abschnitt 4.1 beschriebenen Ansatz nach Chen das Gewicht einer jeden Phase mit der Anzahl der Ausführungen der Abfrage d innerhalb dieser Phase ($u_i(d)$) multipliziert und über die Betrachtungsperiode hinweg aufsummiert:

$$v_t(d) = \sum_{i=1}^{n} (w_i \cdot u_i(d))$$

Ein Beispiel soll die Berechnung des Wertes einer Abfrage verdeutlichen. Wir gehen dabei von einer Betrachtungsperiode mit drei Phasen aus. Abbildung 2 veranschaulicht die Nummerierung und Gewichtung der Phasen:

Abbildung 2: Betrachtungsperiode und Phasen

Für x=2 berechnet sich das Gewicht der ersten Phase wie folgt:

$$w_1 = \frac{\left(\frac{1}{2}\right)^1}{\sum_{j=1}^{3}\left(\frac{1}{2}\right)^j} = \frac{0{,}5}{\left(\frac{1}{2}\right)^1 + \left(\frac{1}{2}\right)^2 + \left(\frac{1}{2}\right)^3} = \frac{0{,}5}{0{,}875} = 0{,}5714$$

Auf dem von uns beispielhaft betrachteten Datenwürfel sind zwei Abfragen definiert. Tabelle 1 zeigt die Ausführungsstatistik für die zwei Abfragen in jeder Phase der Betrachtungsperiode.

	Phase 3	Phase 2	Phase 1
Abfrage A1	39	23	15
Abfrage A2	20	12	19

Tabelle 1: Ausführungsstatistik Beispiel

Der Wert für Abfrage A1 ergibt sich aus folgender Berechnung:

$$v_t(A1) = \sum_{i=1}^{3} \left(w_i \cdot u_i(A1) \right) = 0{,}5714 \cdot 15 + 0{,}2857 \cdot 23 + 0{,}1429 \cdot 39 = 20{,}7152$$

Damit im Rahmen von ILM geeignete Regeln für die Klassifizierung der Informationen anhand ihres Nutzungsgrades aufgestellt werden können, sollten die Werte der Abfragen ein fest vorgegebenes Intervall (z. B. [0,10]) nicht verlassen. Um dies zu erreichen, müssen die Werte der Abfragen mit den vorzugebenden Intervallgrenzen normalisiert werden.

Nachdem für alle Abfragen des betrachteten Datenwürfels jeweils ein Wert bestimmt worden ist, wird dieser auf alle Informationen, die eine Abfrage angesprochen hat, übertragen. Werden Informationen von mehreren Abfragen angesprochen, ist den Informationen jeweils der höchste Wert zuzuordnen. Im Gegensatz zur normalen Durchschnittsbildung, bei der alle Abfragen gleichermaßen zur Bildung des Wertes für eine Information beitragen, kann diese Methode nicht zu einer Unterbewertung von Informationen führen. Eine gewichtete Durchschnittsbildung wäre eine Alternative dazu, ist aber aufgrund der zu bestimmenden Gewichte für die einzelnen Abfragen sehr viel schwerer automatisierbar.

Bei der Anwendung des vorgeschlagenen Bewertungsansatzes ist zu prüfen, ob innerhalb der eingegrenzten Betrachtungsperiode neue Abfragen entstanden sind oder gelöscht wurden. Treten solche Fälle auf, müssen die entsprechenden Abfragen gesondert behandelt werden, indem nur die Phasen zur Bestimmung des Wertes berücksichtigt werden, in denen die Abfrage tatsächlich existierte. Zudem sollten speziell für diese Abfragen neue Gewichte für die zu berücksichtigenden Phasen berechnet werden. Die neuen Gewichte müssen in ihrer Summe wieder 1 ergeben. Die Proportionalität zu den ursprünglichen Gewichten muss gewahrt bleiben. Um neben der Entstehung und Löschung von Abfragen auch Modifikationen mit dem vorgeschlagenen Bewertungsansatz abbilden zu können, ist die Einführung einer Versionsverwaltung für Abfragen grundsätzliche Voraussetzung. Es muss jederzeit bestimmbar sein, welche Informationen eine Abfrage zu einem bestimmten Zeitpunkt angesprochen hat. Die Behandlung von Abfragen, deren Version sich innerhalb der Betrachtungsperiode geändert hat, kann dann analog zur Berücksichtigung von neu entstandenen und gelöschten Abfragen erfolgen, indem eine neue Version wie eine eigenständige Abfrage behandelt wird.

Im Anschluss an die Bewertung der Informationen werden diese entsprechend vordefinierter Regeln in verschiedene Klassen eingeteilt. Diese Regeln sind unter Einbeziehung des Anwender- und Administratorenwissens über das betrachtete Data-Warehouse im Vorfeld der Klassifizierung zu definieren und legen fest, welche Werte der Informationen zu welcher Klasse gehören. Auf Grundlage dieser Klassifizierung der Informationen kann die Verlagerung auf die einzelnen Speicherbereiche erfolgen.[8]

[8] Vgl. Abbildung 1, Phase 4.

4.3 Automatisierte Klassifizierung von Informationen am Beispiel des SAP BI

Die praktische Anwendbarkeit des Konzeptes wurde im Rahmen eines SAP-BI[9]-Projektes überprüft. Dieses Projekt wurde im Jahr 2007 im Rahmen einer Diplomarbeit in Zusammenarbeit mit dem Praxispartner IBM Deutschland in Chemnitz realisiert. Die Überprüfung erfolgte in Form einer prototypischen Implementierung des Bewertungsverfahrens einschließlich einer Klassifizierung der Informationen anhand vordefinierter Regeln.

Eine wesentliche Voraussetzung der Anwendung des Konzeptes besteht darin, dass durch das SAP BI statistische Informationen über die Ausführung von Abfragen aufgezeichnet werden. Das System ist in der Lage zu protokollieren, welche Abfrage zu welchem Zeitpunkt ausgeführt worden ist.[10] Zur Implementierung des vorgeschlagenen Bewertungskonzeptes musste aus den aufgezeichneten Informationen des SAP BI eine geeignete Ausführungsstatistik erzeugt werden. Dabei ermittelt der implementierte Prototyp die Summe der Ausführungen einer Abfrage innerhalb einer Phase, wofür Beginn und Ende der Betrachtungsperiode sowie die Anzahl bzw. Länge der Phasen im Vorfeld festzulegen sind. Des Weiteren muss der Parameter x vorgegeben werden, um die Phasen für die folgende Wertbestimmung entsprechend gewichten zu können.

Nach der Bewertung aller Abfragen eines Datenwürfels und der Normalisierung der Abfrage-Werte im Intervall [0,10] werden im Rahmen der prototypischen Implementierung des Konzeptes die Abfrage-Werte auf alle Informationen der Faktentabelle übertragen, die jeweils von einer Abfrage angesprochen werden. Es kommt die beschriebene Maximum-Methode zum Einsatz. Um die von einer Abfrage bei deren Ausführung angesprochenen Informationen der Faktentabelle zu bestimmen, wurden bei der Implementierung des Verfahrens die Selektionsbedingungen der Abfrage ausgewertet.

Im Anschluss an die Bewertung der Informationen werden diese entsprechend vordefinierter Regeln in verschiedene Klassen eingeteilt. Bei der Implementierung des Prototyps wurde die in Tabelle 2 dargestellte Klassifizierung gewählt[11]:

Wert der Information	Klasse
10 – 7	Online
6 – 4	Nearline
3 – 1	Offline
0	Löschen

Tabelle 2: Beispielhafte Klassifizierung

[9] Hierbei handelt es sich um das von der SAP AG vertriebene Produkt SAP NetWeaver Business Intelligence in der Version 7.0.

[10] Hierfür stellt die SAP AG vorgefertigte Objekte innerhalb des so genannten SAP Business Intelligence Content zur Verfügung.

[11] Vgl. die unterschiedlichen Klassen, Abbildung 1, Phase 4.

4.4 Auswertung der Klassifizierungsergebnisse

Untersuchungen auf einem künstlich erzeugten Informationsbestand ergaben, dass durch die Implementierung des vorgeschlagenen Klassifizierungskonzeptes bis zu fünfzig Prozent des Informationsbestandes eines Datenwürfels aus dem Online-Speicherbereich ausgelagert werden kann. Grundlage des künstlich erzeugten Informationsbestandes ist ein Datenwürfel mit 50.095 Informationen. Bei diesen Informationen handelt es sich um anonymisierte Finanzinformationen eines Unternehmens, die von dem Praxispartner IBM Deutschland für die Zwecke der Diplomarbeit zur Verfügung gestellt wurden. Auf dem untersuchten Datenwürfel wurden elf Abfragen mit willkürlich festgelegten Selektionsbedingungen angelegt. Diese Abfragen wurden nach einem definierten Ausführungsplan auf den Datenwürfel angewandt, um möglichst unterschiedliche Ausführungsmuster zu erzeugen und deren Auswirkungen auf die Bewertung der Informationen abschätzen zu können. Untersuchungen mit den von Chen vorgeschlagenen Standardwerten für den Parameter x und die Anzahl bzw. Länge der Phasen ergaben, dass etwa ein Drittel der Informationen der Faktentabelle in den Nearline-Bereich und nahezu ein Sechstel in den Offline-Bereich verschoben werden können. Weniger als ein Prozent der Informationen wurden zur Löschung vorgeschlagen.

Weiterführende Untersuchungen bezüglich der Eingabewerte für den Parameter x und der Anzahl bzw. Länge der Phasen ergaben, dass bei einer Verringerung des Standardwertes für x von 2 auf 1,2 die Abfragen tendenziell höher bewertet werden. Abbildung 3 verdeutlicht dies.

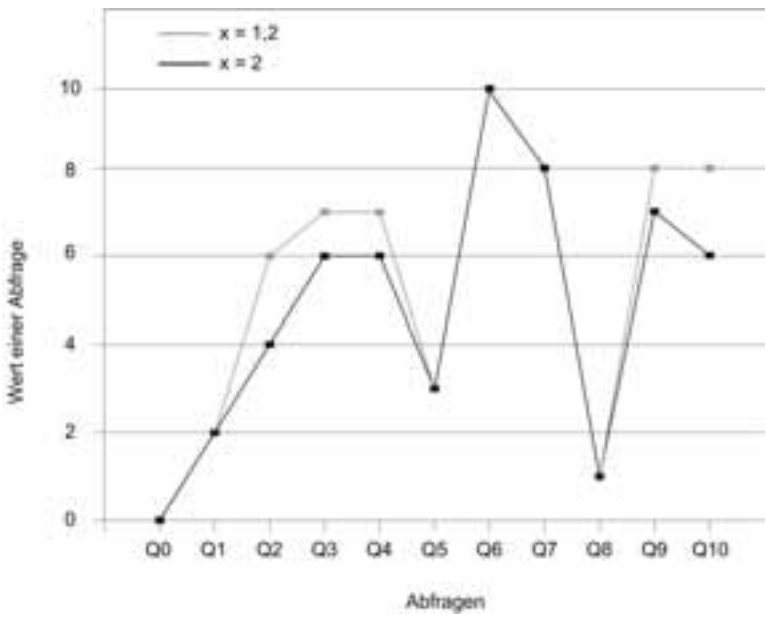

Abbildung 3: Steigende Abfrage-Werte bei sinkendem Parameter x

Der Grund dafür liegt darin, dass sich mit sinkendem Parameter x die Gewichte der Phasen gegenseitig annähern und der Einfluss der Phasen auf die Bewertung der Abfragen sinkt. Die Abfragen werden in diesem Fall tendenziell stärker an der Häufigkeit ihrer Ausführung als an den Ausführungszeitpunkten bewertet, da die sehr flache Gewichtsverteilung der Phasen nicht mehr für eine Abwertung der Ausführungen aus weiter zurückliegenden Phasen sorgt. Dies führt bei dem gegebenen Informationsbestand dazu, dass nur noch ein Sechstel der Informationen der Faktentabelle aus dem Online-Speicherbereich zur Auslagerung vorgeschlagen wird. Je höher der Parameter x ist, desto mehr werden die Ausführungshäufigkeiten bei der Bewertung der Abfragen vernachlässigt. Die Bewertung erfolgt dann tendenziell nach dem Zugriffszeitpunkt.

Gleiche Untersuchungsergebnisse ergaben sich bei Veränderung der Anzahl bzw. Länge der Phasen. Bei der Bewertung der Abfragen werden die Ausführungszeitpunkte tendenziell vernachlässigt, sobald bei einer Betrachtungsperiode von acht Wochen die Anzahl der Phasen kleiner als fünf ist. Wie bereits von Chen beschrieben, ist der Parameter x und die Anzahl bzw. Länge der Phasen so zu wählen, dass Ausführungshäufigkeit und -zeitpunkte gleichermaßen an der Bestimmung der Abfrage-Werte beteiligt sind.

4.5 Evaluierung des Klassifizierungskonzeptes

Es soll nun überprüft werden, inwiefern das von uns entwickelte Konzept die in Abschnitt 3 definierten Anforderungen erfüllt.

Die Kriterien Zugriffshäufigkeit und -zeitpunkte bei der Bestimmung des Nutzungsgrades wurden analog zum Ansatz von Chen verwendet. Damit kann überprüft werden, wie oft und wann auf die zu einem Datenwürfel erstellten Abfragen zugegriffen wird bzw. welche Informationen gelesen werden.

Mit der automatisierten Bewertung und Klassifizierung der Informationen ist eine weitere Anforderung erfüllt. Die Berücksichtigung des Anwender- und Administratorenwissens erfolgt bei der Bestimmung der Betrachtungsperiode, der Anzahl bzw. Länge der Phasen, der Festlegung des Parameters x sowie bei der Aufstellung der Regeln für die Klassifizierung der Informationen. Durch diese Möglichkeiten der Einflussnahme in den Prozess der Informationsklassifizierung ist das Verfahren an die Nutzungsintensität des betrachteten Data-Warehouse-Systems flexibel anpassbar. Eine manuelle Veränderung der Klassifizierung im Anschluss an die automatisierte Klassifikation ist in unserem Konzept nicht möglich.

Die Prognose zur Ermittlung zukünftiger Zugriffshäufigkeiten und -zeitpunkte sowie die Anforderung zur Berücksichtigung rechtlicher Restriktionen erfüllt unser Konzept nicht. Die Informationen werden ausschließlich anhand ihrer Nutzung auf Grundlage von Vergangenheitswerten bewertet. Neben unserem Konzept erfüllt auch keines der in Abschnitt 3 vorgestellten Konzepte zur automatisierten Klassifizierung von Informationen die Berücksichtigung rechtlicher Anforderungen. Dieses Gebiet ist ein offenes Forschungsthema [Ab06], [MB07].

Die Berücksichtigung des zukünftigen Nutzungsgrades ließe sich in unser Bewertungsverfahren integrieren, indem die Betrachtungsperiode, die der Bewertung zugrunde liegt, in die Zukunft ausgedehnt wird. Chen schlägt vor, dass die Gewichtung der Phasen dann in Form einer Glockenkurve erfolgt, die ihren Höhepunkt zur gegenwärtigen Phase erreicht. Zukünftige Phasen erhalten analog zur Vergangenheit sinkende Gewichte, da der Nutzungsgrad der Informationen nur geschätzt und daher unsicher ist. Zur Konkretisierung dieses Vorgehens und dessen Integration in unser Klassifizierungskonzept ist weitere Forschungs- und Entwicklungsarbeit notwendig.

Ein wesentlicher Kritikpunkt an dem von uns vorgeschlagenen Klassifizierungskonzept ist, dass Informationen, die in den Offline-Bereich verlagert wurden, nicht mehr direkt auswertbar sind. Sobald Informationen aber aus dem Bereich der direkten Auswertbarkeit ausgelagert werden, ist die Interpretation der verbleibenden Informationen ohne entsprechende Kennzeichnung sehr schwierig, da beispielsweise Aggregationen nicht mehr vollständig sind. Wie eine solche Kennzeichnung genau aussehen kann, ist ein weiterer Punkt, auf den im Rahmen weiterer Arbeiten näher eingegangen werden muss [St01].

Unser Klassifizierungskonzept behandelt ausschließlich die Klassifikation von Informationen in Datenwürfeln, nicht aber in physischen Aggregaten. Die Berücksichtigung von Aggregaten ist nicht notwendig, da diese ohnehin einer statistischen Auswertung durch das eingesetzte Data-Warehouse-System unterliegen und das System selbst Vorschläge zur Verwendung der in den Aggregaten befindlichen Informationen unterbreitet.[12] Aggregate ermöglichen einen schnellen Zugriff auf die Informationen eines Datenwürfels. Dabei werden die Informationen des Datenwürfels, auf welche innerhalb von Berichten häufig zugegriffen wird, redundant gespeichert, um eine höhere Performanz zu erzielen. Aus diesem Grund sollten die Informationen der Aggregate ausschließlich innerhalb der Speicherklasse „Online“ vorgehalten werden.[13] Werden Informationen in Datenwürfeln verändert, so werden diese durch das so genannte „Hochrollen“ der Informationen in die zugehörigen Aggregate übertragen. Damit wirken sich Auslagerungen von Informationen aus den Datenwürfeln auch automatisch auf alle damit verbundenen physischen Aggregate aus.

Das Potential unseres Klassifizierungskonzeptes wurde seitens des Praxispartners IBM Deutschland erkannt und wertgeschätzt, so dass derzeit an der Weiterentwicklung der prototypischen Realisierung in SAP BI gearbeitet wird. Als hauptsächliche Herausforderung stellt sich hierbei der Ausbau des Vorgehens zur Bestimmung der von einer Abfrage angesprochenen Informationen. Neben statisch definierten Selektionsbedingungen sollen nun auch variable, während der Abfrageausführung festgelegte Selektionsbedingungen ausgewertet werden. Das sich hierbei ergebende Problem besteht darin, dass in SAP BI die dynamischen Selektionsbedingungen nur protokolliert werden, sofern zur Ausführung der Abfrage auf die Datenbank zurückgegriffen werden muss. Erfolgt das

[12] Detaillierte Aussagen zur Verwendung bzw. zu den Zugriffen auf die Informationen von Aggregaten beispielsweise in SAP BI liefert die BI-Statistik.

[13] Vgl. Abschnitt 2

Lesen der Abfrageergebnisse jedoch vollständig aus dem Zwischenspeicher des Systems, findet keine Aufzeichnung der verwendeten Selektionsbedingungen statt. Hier weist das SAP BI bezüglich der Realisierung unseres Konzeptes ein klares Defizit auf.

5 Zusammenfassung und Ausblick

Unser Beitrag konzentriert sich auf die zentrale Teilaufgabe des ILM, die Informationsklassifizierung. Nach der Einordnung dieser Teilaufgabe in den Gesamtkontext des ILM im Data-Warehouse-Bereich haben wir auf Grundlage der Vorarbeiten von Chen ein Konzept vorgestellt, welches Informationen in einem Data Warehouse entsprechend ihres Nutzungsgrades automatisiert klassifiziert. Anhand einer prototypischen Implementierung im Rahmen eines SAP-BI-Projektes konnten wir die Praktikabilität des Konzeptes demonstrieren und Verbesserungsmöglichkeiten aufzeigen.

Der Vorteil unseres Klassifikationskonzepts besteht in der automatisierten Ermittlung der Zugriffshäufigkeiten auf einzelne Informationen in Datenwürfeln. Dieses Verfahren wurde speziell für das Data-Warehouse-Umfeld entwickelt, wodurch unser Konzept nicht ohne Weiteres auf OLTP-Systeme übertragen werden kann.

Im Fokus der zukünftigen Forschung und Entwicklung steht die Weiterentwicklung von ILM-Software [GM05]. Eine Software, die automatisiert Vorschläge bezüglich des Wertes bzw. der Klassifikation von Informationen liefert, ist für die Entwicklung und Umsetzung von ILM-Konzepten im Data-Warehouse-Umfeld von Vorteil, wenngleich Eingriffsmöglichkeiten der Verantwortlichen bzw. Anwender weiterhin notwendig sind.

6 Literaturverzeichnis

[Ab05] Abd-El-Malek, M. et al.: Ursa Minor: versatile cluster-based storage. In (Gibson G. Hrsg.): Proceedings of the 4th USENIX Conference on File and Storage Technologies, San Francisco, 2005; S. 59-72.

[Ab06] Abd-El-Malek, M. et al.: Early experiences on the journey towards self-* storage. In (Lomet, B. Hrsg.): Data Engineering Bulletin, Ausgabe 29, Nr. 4, Atlanta, 2006; S. 55-62.

[Al01] Allen, N.: Don't waste your storage dollars: what you need to know. Research Note, Gartner Group Inc., Stamford 2001.

[Bh05] Bhagwan R. et al.: Time-varying Management of Data Storage. In (Candea G.; Oppenheimer, D. Hrsg.): First Workshop on Hot Topics in Systems Dependability, Yokohama, 2005.

[CGH05] Chamoni, P.; Gluchowski, P.; Hahne, M.: Business Information Warehouse: Perspektiven betrieblicher Informationsversorgung und Entscheidungsunterstützung auf der Basis von SAP Systemen. Berlin, Heidelberg, New York, 2005; S. 36-39.

[CGY07] Chandra, S.; Gehani, A.; Yu, X.: Automated Storage Reclamation Using Temporal Importance Annotations. In: 27th International Conference on Distributed Computing Systems, Toronto, 2007; S. 12.

[Ch05] Chen Y.: Information valuation for Information Lifecycle Management. In (Parashar, M. et al. Hrsg.): Proceedings of the 2nd International Conference on Autonomic Computing, Seattle, Washington, 2005; S. 135-146.

[Do04] Douglis, F. et al.: Position: Short Object Lifetimes Require a Delete-Optimized Storage System. In: Proceedings of the 11th workshop on ACM SIGOPS European workshop: beyond the PC, ACM Press, New York, 2004.

[El03] Ellard, D. et al.: Attribute-Based File Prediction of File Properties. Harvard, 2003; S. 1-14.

[GM05] Gillet, F.; Mendel, T.: Organic IT: IT-Kosten senken, Unternehmensabläufe beschleunigen. In: (Kuhlin, B.; Thielmann, H. Hrsg.): Real-time Enterprise in der Praxis: Fakten und Praxis. Berlin, Heidelberg, New York, 2005; S. 483-502.

[Gr03] Gray, J.: A conversation with Jim Gray. In (ACM Press Hrsg.): Queue Storage, Ausgabe 1, Nr. 4, New York, 2003; S. 8-17.

[HS96] Hillyer, B.; Silberschatz, A.: Random I/O scheduling in online tertiary storage systems. In: ACM SIGMOD International Conference on Management of Data, 1996; S. 195-204.

[IIS01] Inmon, W.; Imhoff, C.; Sousa, R.: Corporate Information Factory. 2. Auflage, New York et al. 2001; S. 139-156.

[Ja98] Janko, W.: Informationswirtschaft 1. Grundlagen der Informatik für die Informationswirtschaft. Springer-Verlag, Berlin, 1998.

[KSH96] Küspert, K.; Schaarschmidt, R.; Herbst, A.: Archiv Transaktionen: Ein Ansatz für asynchrones, transaktionsgesichertes Archivieren großer Datenmengen in Datenbanksystemen. In: ITG-Fachbericht zur 4. ITG/GI/GMA-Fachtagung Softwaretechnik in Automation und Kommunikation – Rechnergestützte Teamarbeit (STAK). München, 1996; S. 195-211.

[LLZ04] Liu, B.; Li, J.; Zhang, Y.: Optimal Data Dispatching Methods in Near-Line Tertiary Storage System. In: (Li, Q.; Wang, G.; Feng L. Hrsg.): Proceedings of the 5th International Conference: Advances in Web-Age Information Management. Dalian, China, 2004, S. 690-695.

[MB07] Mont, M.; Beato, F.: On Parametric Obligation Policies: Enabling Privacy-aware Information Lifecycle Management in Enterprises. In (IEEE Computer Society): Proceedings of the 8th IEEE International Workshop on Policies for Distributed Systems and Networks, Washington, 2007; S. 51-55.

[Me04] Mesnier, M. et al.: File classification in self-* storage systems. In (Ibrahim, M. Hrsg.): Proceedings of the 1st International Conference on Autonomic Computing (ICAC-04). New York 2004; S. 44-51.

[MHP05] Maier, R.; Hädrich, T.; Peinl, R.: Enterprise Knowledge Infrastructures. Springer-Verlag, Berlin, Heidelberg, New York 2005; S. 109, 116, 254-257.

[Mo06] Mont, M.: On Privacy-aware Information Lifecycle Management in Enterprises: Setting the Context. In (Paulus, S.; Pohlmann, N.; Reimer, H. Hrsg.): ISSE 2006 – Securing Electronic Business Processes. Vieweg-Verlag, Wiesbaden 2006; S. 405-415.

[MS08] Matthesius, M.; Stelzer, D.: Analyse und Vergleich von Konzepten zur automatisierten Informationsbewertung im Information Lifecycle Management. In (Bichler, K. et al. Hrsg.): Multikonferenz Wirtschaftsinformatik 2008. München, 2008; S. 471-482.

[MW99] Moody, D.; Walsh, P.: Measuring the value of information: An asset valuation approach. In (Heje, J. P. et al. Hrsg.): Proceedings of the 7th European Conference on Information Systems. Copenhagen, Frederiksberg, 1999; S. 496-512.

[Pa05] Patrascu, R. et al.: New approaches to optimization and utility elicitation in autonomic computing. Technical Report, Toronto, 2005.

[Pe05] Petrocelli, T.: Data Protection and Information Lifecycle Management. Prentice Hall International, New York et al., 2005; S. 23.

[Sha06] Shah, G. et al.: ACE: Classification for Information Lifecycle Management. Almaden et al., 2006; S. 1-15.

[Sh06] Short, J.: Information Lifecycle Management: An Analysis of End User Perspectives. San Diego, 2006; S. 2-37.

[St01] Stock, S.: Modellierung zeitbezogener Daten im Data Warehouse. Zugleich Dissertation an der Universität Duisburg 2000, Wiesbaden 2001; S. 133.

[TS07] Thome, G.; Sollbach W.: Grundlagen und Modelle des Information Lifecycle Management. Springer-Verlag, Berlin, Heidelberg, New York, 2007; S. 22-30, 152, 171-172.

[Ve05] Verma, A. et al.: An Architecture for Lifecycle Management in Very Large File Systems. In: 13th NASA Goddard, 22nd IEEE Conference on Mass Storage Systems and Technologies (MSST 2005), Monterey, 2005; S. 160-168.

[WW07] Wixom, B.; Watson, H.: Introduction to the Minitrack on Data Warehousing and Business Intelligence. In: Proceedings of the 40th Hawaii International Conference on System Sciences, 2007; S. 214.

[YB02] Yates-Mercer, P.; Bawden, D.: Managing the paradox: the valuation of knowledge and knowledge management. In: Journal of Information Science, Ausgabe 28, Nr.1, London, 2002; S. 19-29.

[Za04] Zadok, E. et al.: Reducing Storage Management Costs via Informed User-Based Policies. In: 12th NASA Goddard, 21st IEEE Conference on Mass Storage Systems and Technologies (MSST 2004), Maryland, 2004; S. 101-105.