Jan 17th, 12:00 AM

Methoden zur Identifikation relevanter Datenquellen: Eine Literaturanalyse

Majeed Khan Malik
Hochschule Heilbronn, Germany, mmalik1@stud.hs-heilbronn.de

Helmut Beckmann
Hochschule Heilbronn, Germany, helmut.beckmann@hs-heilbronn.de

Methoden zur Identifikation relevanter Datenquellen: Eine Literaturanalyse

Majeed Khan Malik[1] und Helmut Beckmann[1]

[1] Hochschule Heilbronn, Wirtschaftsinformatik, Heilbronn, Deutschland
{mmalik1,helmut.beckmann}@hs-heilbronn.de

Abstract. Die zunehmende Anzahl an zu verarbeitenden Unternehmensdaten und die steigende strategische Relevanz des Informationsmanagements formulieren einen Bedarf zur Systematisierung des Prozesses zur Identifikation und Selektion relevanter Datenquellen. Obwohl die Datenquellenauswahl eine zentrale Aufgabe im Management der Informationswirtschaft (im Rahmen des strategischen Informationsmanagements) darstellt, fehlt es einer aktuellen Betrachtung zum Stand der Wissenschaft im Hinblick vorhandener Methoden zur Selektion relevanter Datenquellen.
Der vorliegende Beitrag schließt diese Forschungslücke und stellt den aktuellen Stand der Wissenschaft zu vorhandenen Methoden zur Identifikation und Selektion relevanter Datenquellen dar. Mittels einer Literaturanalyse wurden insgesamt 37 wissenschaftliche Beiträge identifiziert, welche acht Methoden zur Datenquellenauswahl beschreiben. Die identifizierten Methoden wurden anschließend den Kategorien automatisierte, semi-automatisierte und manuelle Verfahren zugeordnet. Dabei konnte ebenfalls eine zunehmende Tendenz zu automatisierten Methoden zur Datenquellenauswahl beobachtet werden.

Keywords: Datenquellenauswahl, Datenmanagement, Literaturanalyse, Informationsmanagement

1 Einleitung

Durch die zunehmende Digitalisierung und unternehmensübergreifenden Kommunikation wird die Anzahl an zu verarbeitenden Unternehmensdaten weiter exponentiell steigen [1]. Die Autoren [2] beschreiben, dass in der Vergangenheit die Verfügbarkeit von Daten ein zentrales Problem für Unternehmen darstellte, heute sei jedoch vielmehr die Identifikation und Selektion relevanter Datenquellen eine Herausforderung. Aufgrund der zunehmenden Bedeutung von Daten als strategische Ressource und vierten Produktionsfaktor [3] stellt die Identifikation relevanter Datenquellen nach [4] eine zentrale Aufgabe eines Unternehmens im Rahmen des Informationsmanagements dar. Die Identifikation relevanter Datenquellen bedarf nach [5] geeigneter Methoden und Werkzeuge, weshalb ein Überblick zu vorhandenen Methoden zur Identifikation relevanter Datenquellen notwendig ist.

Das Ziel der vorliegenden Arbeit ist es, mittels einer systematischen Literaturanalyse Methoden zur Identifikation relevanter Datenquellen zu identifizieren und deren

17th International Conference on Wirtschaftsinformatik,
February 2022, Nürnberg, Germany

Funktionsweise zu beschreiben. Durch diese Arbeit soll somit der aktuelle Stand der Wissenschaft an Methoden zur Identifikation relevanter Datenquellen im Rahmen des strategischen Informationsmanagements dargestellt werden. Die vorliegende Arbeit leistet einen Beitrag zum Überblicken vorhandener Methoden zur Selektion relevanter Datenquellen und dient als Orientierungshilfe für weitere Forschungsvorhaben in diesem Fachgebiet.

Im Rahmen dieser Arbeit werden folgende zentrale Forschungsfragen beantwortet:

Q1: Welche zentralen Methoden lassen sich zur systematischen Selektion relevanter Datenquellen identifizieren?
Q1.1: Wie können die identifizierten Methoden klassifiziert werden?
Q1.2: Kann aus den identifizierten Methoden ein Trend hinsichtlich bestimmter Technologien oder Methoden erkannt werden?

Die vorliegende Arbeit ist wie folgt aufgebaut: Im nachfolgenden Abschnitt werden zentrale Begriffe zum Verständnis dieser Arbeit definiert. Anschließend erfolgt die Darstellung der Methodik zur Literaturanalyse in Abschnitt 3. In Abschnitt 4 werden die durch die durchgeführte Literaturanalyse gewonnenen Forschungsergebnisse diskutiert und die Forschungsfrage beantwortet. Abschließend erfolgt in Abschnitt 5 eine Zusammenfassung der Erkenntnisse, Diskussion und kritische Würdigung der durchgeführten Forschungsarbeit und ein Ausblick für zukünftige Forschungsvorhaben.

2 Grundlagen und Begriffe

2.1 Data source selection

In der wissenschaftlichen Literatur wird die Aufgabe zur Selektion relevanter Datenquellen als „data source selection“ [6] bezeichnet. Dieser Begriff wird im Rahmen von dieser Arbeit verwendet und wird nachfolgend definiert. Nach [6] bezeichnet man unter data source selection die Aufgabe zur Datenquellenauswahl aus einer Gesamtmenge $\Omega = \{S_1, S_2, \ldots, S_n\}$ bestehend aus n Datenquellen. Dabei ist jede Datenquelle S_i durch entsprechende Kosten C_i, Abdeckung Cov_i, Mehrwert G_i und Qualität Q_i definiert. Die Aufgabe besteht darin, im Hinblick dieser Charakteristiken eine Teilmenge $S^* \subseteq \Omega$ an Datenquellen auszuwählen, wobei der Informationsmehrwert (Gain) maximiert und die Kosten nicht überschritten werden [6]. [7, S.130] bezeichnet diese Aufgabe als *Management der Informationsquellen*, welches das Erkennen, Erheben, Sammeln und Erfassen von Informationsquellen zum Ziel hat.

2.2 Einordnung der Datenquellenauswahl in das Informationsmanagement

Die Aufgabe zur Selektion relevanter Datenquellen lässt sich nach dem Modell des Informationsmanagements von [7] in die Disziplin *Management der*

Informationswirtschaft einordnen. Dies hat nach [7, S.107] das „*Treffen von Entscheidungen über den Informationsbedarf und das Informationsangebot, damit um den Informationseinsatz*“ zur Aufgabe. [8, S.236] nennt in diesem Rahmen, dass die unternehmerische Aufgabe darin besteht, die Informationsnachfrage durch ein Informationsangebot auf allen Unternehmensbereichen zu decken. In diesem Kontext kommt der Identifikation relevanter Datenquellen eine entscheidende Bedeutung zu, da dies den Rahmen für das Informationsangebot definiert.

2.3 Relevante Datenquellen

Die Autoren [4] beschreiben, dass nicht pauschalisiert werden kann, welche Datenquellen relevant sind oder nicht. [4] beschreiben, dass die Auswahl relevanter Datenquellen ausschließlich in einem definierten Kontext (z. B. im Rahmen einer Entscheidungsfrage) getroffen werden kann. [9] nennen hierbei den Begriff „*fitness for Purpose*“, demnach kann die Auswahl relevanter Datenquellen ausschließlich im Kontext des Einsatzzweckes beantwortet werden.

3 Methodik zur Literaturanalyse

Zur Beantwortung der Forschungsfrage wird eine fünfstufige Literaturanalyse und -review nach [10] durchgeführt. Durch die Anwendung dieser Methodik können relevante wissenschaftliche Publikationen identifiziert und Forschungsergebnisse belastbar nachvollzogen werden.

Im ersten Schritt der Literaturanalyse nach [10] wurde die Forschungsfrage inkl. Problemstellung beschrieben und zu anderen Forschungstätigkeiten abgegrenzt. Anschließend erfolgt im zweiten Schritt nach [10] die Literatursuche. In diesem Beitrag erfolgt die Literatursuche nach [11], da mittels dieser Methodik eine möglichst hohe Abdeckung an relevanter Literatur erzielt werden kann. In der Literatursuche werden zunächst Suchterme (deutschsprachig und englischsprachig) spezifiziert, welche zur Identifikation relevanter Literatur verwendet werden. Nachfolgend ist eine Auswahl relevanter Suchterme dargestellt:

Deutschsprachig:

- „datenquelle“ AND (auswahl OR strategie OR methode OR technik)
- informationswirtschaft AND („relevante datenquellen“ OR „informationsquelle“)
- ("selektion relevanter datenquellen" OR "datenquellenauswahl") AND (methode OR technik OR verfahren OR strategie)

Englischsprachig:

- „data source selection” AND (strategy OR method OR technique)
- „information lifecycle“ AND („relevant data“ OR „datasource“)
- "data source selection” AND (categorization OR classification OR grouping OR organization)

Die Literatursuche wurde in definierten Literaturdatenbanken (*AIS Electronic Library, Association for Computing Machinery Digital Library, EmeraldInsight, Google Scholar, IEEE Xplore Digital Library, ScienceDirect*) und ausgewählten Tagungsbänder von Konferenzen (*International Conference on Data Quality Management, International Conference on Information Quality, Internationale Tagung Wirtschaftsinformatik, International Conference on Information Reuse and Integration for Data Science*) durchgeführt. Zur weiteren Eingrenzung der Suchergebnisse wird die Suche auf die Felder Titel, Abstract und Keywords, sowohl auf den Zeitraum 2010 – 2021 begrenzt.

Durch diese Vorgehensweise konnten insgesamt 315 Beiträge identifiziert werden. [10] beschreibt, dass im dritten Schritt die Literaturauswertung hinsichtlich der Relevanz erfolgt. Durch die Entfernung von Duplikaten und Sichtung des Abstracts konnte die Anzahl relevanter Publikationen auf 66 reduziert werden. Durch eine anschließende Überprüfung auf die wissenschaftliche Eignung der Beiträge konnten 2 weitere Quellen entfernt werden. Mittels einer Vor- und Rückwärtssuche konnten vier weitere Arbeiten identifiziert und durch eine abschließende inhaltliche Analyse die Gesamtanzahl relevanter Beiträge im Rahmen dieser Arbeit auf 37 reduziert werden (ausgeschlossen von Grundlagenliteratur und Literatur zur Forschungsmethodik). Diese Beiträge wurden anschließend nach ihren diskutierten Konzepten nach [11] kategorisiert.

Nach [10] werden anschließend diese identifizierten Beiträge auf die definierte Problemstellung und Forschungsfrage untersucht. Im fünften und letzten Schritt der Literaturanalyse nach [10] erfolgt die Präsentation der Forschungsergebnisse in Form von diesem Beitrag.

4 Forschungsergebnisse

4.1 Kriterien zur Selektion von Datenquellen

Die Autoren [9] beschreiben, dass nicht generalisiert werden kann, welche Datenquellen eine Relevanz für ein Unternehmen darstellen. Es muss entsprechend dem Einsatzzweck evaluiert werden, welche Datenquellen die Informationsnachfrage befriedigen („*fitness for purpose*“). Dennoch bedarf es nach [7, S.130f.] qualitativer Kriterien, welche den Prozess zur Selektion relevanter Datenquellen unterstützen.

[6] beschreibt, dass Datenquellen durch *Kosten, Informationsmehrwert* und einer *Abdeckung* definiert sind. Die Kosten stellen dabei die Kosten zum Erwerb und Integration der Datenquellen in die Unternehmensinfrastruktur dar (*Total Cost of Ownership*). Der Informationsmehrwert entspricht nach [6], [12-13] dem Mehrwert, welcher durch das Hinzufügen der Datenquelle im Hinblick auf die bereits existierenden Datenquellen im Unternehmen erreicht wird. Die Autoren [14] nennen ebenfalls die Abdeckung als Kriterium für relevante Datenquellen. Hierbei beschreiben sie, dass die Abdeckung ein Maß dafür ist, wie weit eine Datenquelle die Informationsnachfrage befriedigen kann.

Die Autoren [14-15] nennen die *Aktualität* und *Genauigkeit* als weitere qualitative Kriterien von Datenquellen, welche bei der Selektion berücksichtigt werden sollten. Die Aktualität beschreibt nach [14-15] die zeitliche Relevanz einer Datenquelle und ist somit ein Maß für dessen aktueller Richtigkeit. Die Genauigkeit ist nach [14-15] ebenfalls ein Faktor für die Richtigkeit einer Datenquelle, welche beschreibt, ob die Daten inhaltlich genau sind.

[7, S.132] nennt die Kosten und Aktualität ebenfalls als relevante Kriterien zur Selektion von Datenquellen, erweitert jedoch die *Zugänglichkeit* zu diesen Datenquellen als Kriterium. Demnach muss bei der Selektion von Datenquellen deren Zugänglichkeit berücksichtigt werden, da nicht zugängliche Datenquellen einen geringen Mehrwert für das Unternehmen darstellen.

[16, S.23ff.] beschreiben, dass neben der Qualität von Datenquellen zudem die Informationsqualität (kurz IQ) der tatsächlichen Daten betrachtet werden muss. [16, S.28f.] ordnet die IQ in vier Kriterien (zweckabhängig, systemunterstützt, inhärent und darstellungsbezogen) ein. In der nachfolgenden Abbildung sind diese Dimensionen abgebildet.

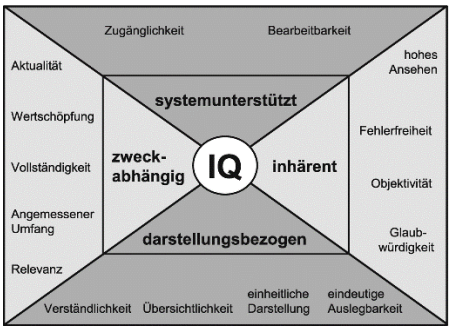

Abbildung 1. IQ Dimensionen nach Kategorie aus [16, S.28f]

Im Rahmen dieser Arbeit werden somit Methoden zur Selektion relevanter Datenquellen betrachtet, welche die zuvor definierten qualitativen Kriterien für Datenquellen und Daten (IQ) berücksichtigen.

4.2 Methoden zur Auswahl von Datenquellen

Nachdem qualitative Kriterien zur Selektion relevanter Datenquellen und zur Informationsqualität beschrieben wurden, werden nachfolgend diese Erkenntnisse angewendet, um Methoden zur Auswahl relevanter Datenquellen zu diskutieren. Die Autoren [17] beschreiben, dass bei Methoden zur Auswahl relevanter Datenquellen zwischen automatischen, semi-automatischen und manuellen Methoden unterschieden werden kann. Diese Kategorisierung wird nachfolgend verwendet, um die identifizierten Methoden zu klassifizieren (Forschungsfrage Q1.1).

4.3 Automatische Ansätze

Zunächst werden automatisierte Methoden zur Selektion relevanter Datenquellen beschrieben. Nach [17] handelt es sich hierbei um Methoden, welche automatisiert relevante Datenquellen im Hinblick einer Frage identifizieren, ohne dass Experten im Auswahlprozess notwendig sind.

Machine Learning Verfahren

Die Autoren [18] beschreiben, dass aufgrund der zunehmenden Datenanzahl eine manuelle Verarbeitung von Datenquellen durch Fachexperten nicht mehr möglich sei, weshalb es einer automatisierten Verarbeitung bedarf. [19] beschreiben einen *Topic Modeling Ansatz* auf Basis des Latent Dirichlet Allocation (kurz LDA) Modell zur Identifikation zentraler Themen in textuellen Datenquellen. Mittels dieses Ansatzes sollen Themen in Dokumenten identifiziert werden, sodass evaluiert werden kann, ob diese im Hinblick einer Entscheidungsfrage relevant sind. [18] beschreiben ebenfalls einen Topic Modeling Ansatz auf Basis des LDA Modells, erweitern jedoch den Ansatz von [19] um die dezentrale Datenspeicherung in einer Blockchain.

Die Autoren [20] beschreiben ein *unüberwachtes Machine Learning Verfahren*, bei welchem auf Basis einer definierten Abfrage („*purpose*“) potenziell relevante Datenquellen identifiziert werden. Hierbei wird die Menge an zu betrachteten Datenquellen mittels eines Clustering-Algorithmus auf ihre Relevanz untersucht. Dieses Verfahren bezieht ebenfalls die bereits existierenden Datenquellen im Unternehmen ein und kann somit ähnliche bzw. redundante Datenquellen automatisch identifizieren. Somit erfolgt eine automatisierte Bewertung des Qualitätskriteriums Informationsmehrwert. Die Qualitätskriterien Kosten, Aktualität und Zugänglichkeit können durch Regeln sichergestellt werden, die Kriterien Abdeckung und Genauigkeit werden nicht berücksichtigt. Die Autoren [21-22] beschreiben vergleichbare Ansätze und evaluieren dabei weitere Clustering Algorithmen.

Im Vergleich zu [20] beschreiben die Autoren [23] *überwachte Machine Learning Verfahren* zur Identifikation relevanter Datenquellen. Hierbei wird auf Grundlage von Metadaten bereits integrierter Datenquellen ein Machine Learning Modell trainiert, welcher neue Datenquellen automatisiert bewertet. Bei diesem Ansatz ist jedoch zu beachten, dass jene Datenquellen als relevant betrachtet werden, welche ähnliche Metadaten wie die bereits integrierten Datenquellen aufweisen. Daher eignet sich dieses Verfahren nicht zur Identifikation divergenter Datenquellen.

Die Autoren [24] beschreiben ebenfalls ein Machine Learning Modell auf Grundlage der Metadaten einer Datenquelle. Jedoch haben [24] ein unüberwachtes Lernverfahren in Form des Rankings implementiert, bei welchem eine Menge an Datenquellen auf Basis definierter Qualitätskriterien in eine sortierte Reihenfolge gebracht werden. Die Autoren [24] beschreiben, dass die Qualitätskriterien und Einschränkungen der Metadaten durch einen Nutzer definiert werden müssen. Demnach benötigt dieses System als Input Expertenwissen und kann anschließend automatisiert potenziell relevante Datenquellen identifizieren. Die Autoren [25-26] beschreiben ebenfalls ein Machine Learning Ansatz auf Basis eines Ranking-Modells zur Identifikation relevanter Datenquellen. [26] begrenzen sich XML-Dokumente.

Zusammenfassend lässt sich festhalten, dass aktuelle Forschungstätigkeiten die automatisierte Identifikation relevanter Datenquellen mittels unüberwachter Machine Learning Verfahren fokussieren. So konnten acht Publikationen identifiziert werden, welche unüberwachte Machine Learning Verfahren untersuchen, lediglich eine Quelle untersucht überwachte Machine Learning Verfahren.

Algorithmen

Algorithmen sind vergleichbar zu den bereits beschriebenen Machine Learning Verfahren, jedoch verwenden diese keine Form der künstlichen Intelligenz, sondern nutzen u. a. statische Regeln (beispielsweise in Form von Schwellenwerten) und fest definierten Verarbeitungs- bzw. Entscheidungsabläufe [27-29]. Die Autoren [27] haben beispielsweise einen Algorithmus zur automatisierten Ermittlung des Informationsmehrwertes entwickelt und stellen den ermittelten Wert mit den Kosten gegenüber. Mittels dieses Verfahrens können Datenquellen automatisiert auf Basis dieser Qualitätskriterien ausgewählt werden (wenn Informationsmehrwert > Kosten). Die Autoren [28] verfolgen einen vergleichbaren Ansatz, berücksichtigen jedoch ebenfalls die Aktualität der Datenquellen bei der Entscheidungsfindung. Die Autoren [29] und [30] haben je einen Algorithmus entwickelt, welcher den Informationsmehrwert und die Aktualität betrachtet, jedoch nicht die Kosten zum Erwerb und Integration in den Auswahlprozess berücksichtigt.

Quality-driven data source selection

Die Autoren [31] beschreiben mit dem Begriff *quality-driven data source selection (kurz QDDSS)* ein ähnliches Verfahren wie [27] zur automatisierten Identifikation relevanter Datenquellen. Auf Grundlage definierter Qualitätskriterien werden automatisiert relevante Datenquellen identifiziert. Im Gegensatz zu [27], werden ausschließlich qualitative Kriterien bei der Auswahl berücksichtigt.

[32] beschreiben die Implementation eines Abfragesystems, welches auf Basis zuvor definierter Qualitätskriterien, Datenquellen nach ihrer Relevanz bewertet. [32] beschreiben jedoch, dass ihr System nicht für große Datenmengen ausgelegt ist und somit nicht den aktuellen Herausforderungen gerecht wird. Die Autoren [31] beschreiben ein Modell, welches alle zuvor definierten Kriterien bei der Auswahl berücksichtigt. Jedoch beschreiben [31], dass dieses Modell aktuell ausschließlich auf Testdaten evaluiert wurde und eine umfangreiche praktische Evaluierung ausstehe.

Die Autoren [33-34] beschreiben ebenfalls einen qualitativ orientierten Ansatz, berechnen jedoch eine Punktzahl je Datenquelle mittels der Kriterien: (1) Quellähnlichkeit entsprechend dem Benutzerinteresse, (2) Quellähnlichkeit gemäß der Benutzerabfrage („*purpose*“) und (3) der Genauigkeit zwischen den Benutzerpräferenzen und den Merkmalen der Datenquelle. Die Autoren [13] und [35] beschreiben einen qualitativen Ansatz, welcher darauf abzielt, die Gesamtabdeckung durch die Datenquellen zu maximieren, in dem nacheinander Datenquellen nach absteigender Abdeckung integriert werden (bis zu einem Schwellenwert).

Query federation engine

Bei den automatisierten Verfahren zur Selektion relevanter Datenquellen konnten die *Query federation engines* als weitere Methode identifiziert werden. Die Autoren [36] beschreiben dieses Verfahren als Auswahl von relevanten Datenquellen auf Grundlage eines definierten Datenschemas. Dieses Datenschema besteht nach [36] aus folgenden Bestandteilen: (1) Eine interessante Sicht auf eine gegebene Anwendungsdomäne, (2) Quellenpräferenzen und (3) einer Architekturanforderung. Dieses Datenschema wird nach [36] anschließend einem Mediator (Vermittler) übergeben, welcher automatisiert potenziell relevante Datenquellen identifiziert (federation engine). Die Autoren [24] beschreiben ein vergleichbares Verfahren, bei welchem ein Fachexperte eine Abfrage (Query) auf Basis der Metadaten spezifiziert und der Mediator automatisiert potenziell relevante Datenquellen identifiziert. [37] beschreiben ebenfalls ein solches Verfahren, fokussieren sich jedoch ausschließlich auf Webdaten.

Umsetzungshinweise und Voraussetzung

Die automatisierte Selektion relevanter Datenquellen stellt hohe Anforderungen an das Datenmanagement. [38-39] haben ein Reifegradmodell zum Datenmanagement erarbeitet. Die automatisierte Selektion relevanter Datenquellen bedarf demnach die Reifestufe 4: Optimiert. Diese Reifestufe umfasst nach [38-39], dass ein unternehmensweites Standardarchitekturmodell definiert ist, welches zur Verwaltung der Daten, Datenmodelle und Datenbeziehungen verwendet wird. Dies umfasst nach [38-39] zudem die Integration aller Unternehmensdaten, die Spezifikation einer Datenstrategie und die Spezifikation eines übergreifenden Datenmodells. Entsprechend dem Reifegradmodell von [38-39] müssen alle Aufgaben des Informations- und Datenmanagement die Reifestufe optimiert für eine automatisierte Selektion aufweisen. Dies umfasst nach [7, S.107] die Führungsaufgaben IT-Strategie, IT-Governance, IT-Prozesse, IT-Personal, IT-Controlling und IT-Sicherheit.

4.4 Semi-automatisierte Ansätze

[17] beschrieben semi-automatisierte Verfahren als weitere Kategorie zur Identifikation relevanter Datenquellen. Dabei wird nach [17] ebenfalls das Fachwissen in ein System abgebildet, jedoch ist ein Fachexperte als finale kontrollierende Instanz eingebunden.

Assistenzsysteme

[40] argumentieren, dass die Selektion relevanter Datenquellen (auf einem hohen qualitativen Niveau) nicht automatisiert werden kann, weshalb es einem Assistenzsystem bedarf, welches die Fachexperten bei der Auswahl relevanter Datenquellen unterstützt. Die Autoren [14] und [41] beschreiben einen solchen Ansatz, welcher auf Grundlage der qualitativen Kriterien (1) Abdeckung, (2) Genauigkeit und (3) Aktualität eine automatisierte Entscheidungsvorlage einem Fachexperten zur Verfügung stellt, welcher die finale Auswahl an zu integrierenden Datenquellen trifft. Mittels eines solchen Assistenzsystems wird der Fachexperte bei der Selektion

relevanter Datenquellen unterstützt, indem eine Entscheidungsvorlage bereitgestellt wird, jedoch bleibt die finale Verantwortung bei dem Fachexperten.

Der Beitrag von [5] beschreibt einen alternativen Ansatz an Assistenzsystemen. [5] argumentieren, dass eine automatisiert erzeugte Entscheidungsvorlage zwar den Fachexperten bei seiner Tätigkeit unterstützt, jedoch nicht die Erfahrung des Experten berücksichtigen kann. Deshalb beschreiben [5] ein Datenexplorationssystem (IDEaS), welches es einem Fachexperten ermöglicht, Datenquellen im Hinblick definierter Qualitätskriterien zu untersuchen und dadurch eigenständig eine Auswahl an zu integrierenden Datenquellen zu treffen. [17] beschreiben jedoch, dass bei einem solchen System hohe Anforderungen an den Fachexperten gestellt werden, da dieser sowohl die Struktur als auch Semantik der Informationen evaluieren muss.

Quality-driven data source selection

[31] beschreiben, dass bei *QDDSS* Verfahren zwischen automatisierten und semi-automatisierten Methoden unterschieden wird. Nach [31] ist bei den semi-automatisierten Verfahren ein Fachexperte in den Entscheidungsprozess eingebunden. Die Autoren [42] beschreiben ein solches Abfragesystem, welches auf Basis definierter Qualitätskriterien relevante Datenquellen identifiziert. Das durch [42] definierte System liefert einem Experten einen Vorschlag an Datenquellen, welche in die Unternehmensinfrastruktur migriert werden sollen. Dadurch erfolgte keine automatisierte Selektion potenziell relevanter Datenquellen und der Fachexperte stellt eine weitere Instanz zur Qualitätskontrolle dar. Jedoch betrachtet das System von [42] nicht den Inhalt der Daten, weshalb die Qualitätskriterien Abdeckung und Genauigkeit nicht automatisiert evaluiert werden können. Die Autoren [15] beschreiben ebenfalls ein solches System, welches die Kriterien (1) Abdeckung, (2) Genauigkeit und (3) Aktualität berücksichtigt. Das von [15] entwickelt System liefert auf Basis dieser Kriterien eine Entscheidungsgrundlage, welche durch den Fachexperten evaluiert wird.

Umsetzungshinweise und Voraussetzung

Die semi-automatisierte Selektion relevanter Datenquellen stellt ebenfalls hohe Anforderungen an das betriebliche Datenmanagement [38-39]. Nach dem Reifegradmodell von [38-39] sollte ein Unternehmen mindestens die Reifestufe 3: Managed aufweisen. Dies umfasst die zunehmende Standardisierung von Daten, Metadaten, Datenschema, Domänen und Datenmodellen. Nach [38-39] ermöglicht dies die Nutzung und Wiederverwendung von standardisierten Daten. Zudem ist nach [17] zu beachten, dass Fachexperten mit entsprechenden Kompetenzen benötigt werden.

4.5 Manuelle Verfahren

Nach der Kategorisierung von [17] stellen die manuellen Verfahren die abschließende Form von Methoden zur Selektion relevanter Datenquellen dar. [17] beschreiben, dass bei diesen Verfahren die gesamte Verantwortung zur Selektion relevanter Datenquellen auf einen Fachexperten übertragen wird. Die Autoren [43-44] beschreiben ebenfalls, dass bei manuellen Verfahren die gesamte Verantwortung auf einen Fachexperten übertragen wird, welcher entsprechende Kompetenzen zur Verarbeitung und

Bewertung dieser Datenquellen aufweisen muss. [45] beschreiben einen manuellen *QDDSS* Ansatz, bei welchem der Fachexperte auf Grundlage von qualitativen Kriterien eine manuelle Auswahl an relevanten Datenquellen trifft.

Umsetzungshinweise und Voraussetzung
Zur Umsetzung muss beachtet werden, dass die manuelle Selektion von Datenquellen nach [17] hohe Anforderungen an den Fachexperten stellt. Dieser benötigt neben Kompetenzen im Bereich Datenmanagement ebenfalls Kenntnisse zur Datenvisualisierung, -aufbereitung, -verarbeitung und -integration. Hinsichtlich des Reifegradmodelles von [38-39] werden keine spezifischen Anforderungen an das Datenmanagement definiert.

5 Diskussion und Ausblick

Die durchgeführte Forschungsstudie ist nicht ohne Eingrenzungen, welche jedoch Forschungsmöglichkeiten für zukünftige Arbeiten bieten. Zunächst muss kritisch angemerkt werden, dass im Rahmen von diesem Beitrag eine selektive Literaturanalyse durchgeführt wurde. Es konnten nicht sämtlich publizierte wissenschaftliche Beiträge untersucht werden, welche Methoden zur Selektion relevanter Datenquellen diskutieren. Aufgrund dieser Eingrenzung muss beachtet werden, dass im Rahmen dieser Arbeit ausschließlich zentrale Vorgehensweisen betrachtet werden konnten, die Untersuchung sämtlicher Methoden wurde nicht durchgeführt.

Das Ziel dieser Arbeit war die Identifikation zentraler Methoden zur Selektion relevanter Datenquellen. Zusammenfassend lässt sich festhalten, dass durch die durchgeführte Literaturanalyse sieben Methoden (Forschungsfrage Q1) identifiziert werden konnten, welche in drei Kategorien kategorisiert wurden (Forschungsfrage Q1.1). Aus den betrachteten Methoden ist eine Entwicklung zu automatisierten Verfahren erkennbar. Dabei fokussieren die identifizierten Beiträge die Selektion mittels unüberwachter Machine Learning Verfahren (Forschungsfrage Q1.2).

Aus den beschriebenen Eingrenzungen und erarbeiteten Ergebnissen dieser Arbeit wäre es für zukünftige Arbeiten interessant, eine Marktanalyse hinsichtlich vorhandener Systeme zur Selektion relevanter Datenquellen durchzuführen. Mittels dieser Marktanalyse könnte ein konkreter Leitfaden für Unternehmen erstellt werden, welcher eine Orientierungshilfe darstellt und zu einer iterativen Systematisierung des betrieblichen Informations- und Datenmanagement beitragen kann.

Literaturverzeichnis

1. S. Tewes, B. Niestroj, and C. Tewes, Eds., Geschäftsmodelle in die Zukunft denken. Springer Fachmedien Wiesbaden, 2020.
2. M. Dieye, M. F. Zhani, and H. Elbiaze, “On achieving high data availability in heterogeneous cloud storage systems,” Proc. IM 2017 - 2017 IFIP/IEEE Int. Symp. Integr. Netw. Serv. Manag., pp. 326–334, 2017

3. M. Lee et al., “How to Respond to the Fourth Industrial Revolution, or the Second Information Technology Revolution? Dynamic New Combinations between Technology, Market, and Society through Open Innovation,” 2018
4. Y. Lin, H. Wang, J. Li, and H. Gao, “Data source selection for information integration in big data era,” Inf. Sci. (Ny)., vol. 479, pp. 197–213, Apr. 2019
5. A. Bagozi and D. Bianchini, “IDEAaS: Interactive data exploration as-a service,” Proc. - 2019 IEEE World Congr. Serv. Serv. 2019, pp. 345–348, 2019
6. H. M. Safhi, B. Frikh, and B. Ouhbi, “Data Source Selection in Big Data Context,” 2019
7. H. Krcmar, Informationsmanagement. Springer Berlin Heidelberg, 2015.
8. E. Tiemeyer, Handbuch IT-Management: Konzepte, Methoden, Lösungen und Arbeitshilfen für die Praxis. Carl Hanser Verlag GmbH \& Company KG, 2020.
9. F.-B. Mocnik, H. Fan, and A. Zipf, “Data Quality and Fitness for Purpose,” no. May, 2017
10. P. Fettke, “State-of-the-art des state-of-the-art: Eine untersuchung der forschungsmethode ‘review’ innerhalb der wirtschaftsinformatik,” Wirtschaftsinformatik, vol. 48, no. 4, pp. 257–266, 2006
11. J. Webster and R. T. Watson, “Analyzing the Past to Prepare for the Future: Writing a Literature Review.,” MIS Q., vol. 26, no. 2, pp. xiii–xxiii, 2002,
12. X. L. Dong, B. Saha, and D. Srivastava, “Less is More: Selecting Sources Wisely for Integration,” 2012. Accessed: May 09, 2021.
13. M. Salloum, X. L. Dong, D. Srivastava, and V. J. Tsotras, “Online Ordering of Overlapping Data Sources,” Proc. VLDB Endow., vol. 7, no. 3, pp. 133–144, 2013
14. T. Rekatsinas, X. L. Dong, L. Getoor, and D. Srivastava, “Finding quality in quantity: The challenge of discovering valuable sources for integration,” CIDR 2015 - 7th Bienn. Conf. Innov. Data Syst. Res., 2015.
15. T. Rekatsinas, X. L. Dong, and D. Srivastava, “Characterizing and Selecting Fresh Data Sources,” 2014
16. K. Hildebrand, M. Gebauer, H. Hinrichs, and M. Mielke, Eds., Daten- und Informationsqualität. Springer Fachmedien Wiesbaden, 2018.
17. T. Fischer, F. Bakalov, and A. Nauerz, “An Overview of Current Approaches to Mashup Generation,” 2009
18. R. Doku, D. B. Rawat, and C. Liu, “Towards federated learning approach to determine data relevance in big data,” in Proceedings - 2019 IEEE 20th International Conference on Information Reuse and Integration for Data Science, IRI 2019, Jul. 2019, pp. 184–192
19. M. Eickhoff and J. Muntermann, “How to conquer information overload? Supporting financial decisions by identifying relevant conference call topics,” Pacific Asia Conf. Inf. Syst. PACIS 2016 - Proc., 2016.
20. S. El Allali, D. Blank, W. Müller, and A. Henrich, “Image Data Source Selection Using Gaussian Mixture Models,” in Adaptive Multimedia Retrieval: Retrieval, User, and Semantics, 2008, pp. 170–181.
21. M. Eisenhardt, W. Muller, A. Henrich, D. Blank, and S. El Allali, “Clustering-Based Source Selection for Efficient Image Retrieval in Peer-to-Peer Networks,” in Eighth IEEE International Symposium on Multimedia (ISM’06), 2006, pp. 823–830,
22. D. Song, “Deep web data source selection based on subject and probability model,” in Proceedings of 2016 IEEE Advanced Information Management, Communicates, Electronic and Automation Control Conference, IMCEC 2016, Feb. 2017, pp. 1944–1948
23. P. Stanula, A. Ziegenbein, and J. Metternich, “Machine learning algorithms in production: A guideline for efficient data source selection,” Procedia CIRP, 2018, vol. 78, pp. 261–266

24. J. P. Callan, Z. Lu, and W. B. Croft, “Searching Distributed Collections With Inference Networks,” in In Proceedings of the 18th annual international ACM Sigir conference on research and development in infromation retrieval, 1995, pp. 21–28.
25. M. A. Suryanto, E. P. Lim, A. Sun, and R. H. L. Chiang, “Quality-aware collaborative Question Answering: Methods and evaluation,” in Proceedings of the 2nd ACM International Conference on Web Search and Data Mining, WSDM’09, 2009, pp. 142–151,
26. K. Nguyen and J. Cao, “Article Top-K data source selection for keyword queries over multiple XML data sources,” 2012
27. J. Yang and C. Xing, “Data Source Selection Based on an Improved GreedyGenetic Algorithm,” 2019
28. T. Y. Lim, M. A. Al-Betar, and A. T. Khader, “Taming the 0/1 knapsack problem with monogamous pairs genetic algorithm,” Expert Syst. Appl., vol. 54, pp. 241–250, Jul. 2016,
29. X. Zhang, S. Huang, Y. Hu, Y. Zhang, S. Mahadevan, and Y. Deng, “Solving 0-1 knapsack problems based on amoeboid organism algorithm,” 2019,
30. D. Zou, L. Gao, S. Li, and J. Wu, “Solving 0-1 knapsack problem by a novel global harmony search algorithm,” Appl. Soft Comput., vol. 11, pp. 1556–1564, 2011
31. Y. Lin, H. Wang, S. Zhang, J. Li, and H. Gao, “Efficient quality-driven source selection from massive data sources,” J. Syst. Softw., vol. 118, pp. 221–233, Aug. 2016
32. X. L. Dong, L. Berti-Equille, and D. Srivastava, “Truth discovery and copying detection in a dynamic world,” Proc. VLDB Endow., vol. 2, no. 1, pp. 562–573, Aug. 2009
33. S. Kechid and H. Drias, “Personalizing the source selection and the result merging process,” Int. J. Artif. Intell. Tools, vol. 18, no. 2, pp. 331–354, Apr. 2009
34. S. Kechid and H. Drias, “Personalised distributed information retrieval-based agents,” Int. J. Intell. Syst. Technol. Appl., vol. 9, no. 1, pp. 49–74, 2010
35. D. Hong, L. Si, P. Bracke, M. Witt, and T. Juchcinski, “A Joint Probabilistic Classification Model for Resource Selection,” in Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval, 2010, pp. 98–105
36. G. Bruno, C. Collet, and G. Vargas-Solar, “Configuring Intelligent Mediators Using Ontologies,” 2006.
37. M. Goncalves and L. Tineo, “WWW Data Source Selection with SQLfi,” in The 14th IEEE International Conference on Fuzzy Systems, 2005. FUZZ ’05., 2005, pp. 1002–1007
38. K. S. Ryu, J. S. Park, and J. H. Park, “A data quality management maturity model,” ETRI J., vol. 28, no. 2, pp. 191–204, 2006
39. P. Aiken, M. D. Allen, B. Parker, and A. Mattia, “Measuring data management practice maturity: A community’s self-assessment,” Computer, vol. 40, no. 4, pp. 42–50, Apr. 2007
40. S. B. Huffman and D. Steier, “A navigation assistant for data source selection and integration,” 1995.
41. T. Rekatsinas, A. Deshpande, X. L. Dong, L. Getoor, and D. Srivastava, “SourceSight: Enabling effective source selection,” in Proceedings of the ACM SIGMOD International Conference on Management of Data, Jun. 2016, vol. 26-June-20, pp. 2157–2160
42. W. Meng and C. T. Yu, “Advanced Metasearch Engine Technology,” Synth. Lect. Data Manag., vol. 2, no. 1, pp. 1–129, 2010
43. J. Bleiholder, S. Khuller, F. Naumann, L. Raschid, and Y. Wu, “Query planning in the presence of overlapping sources,” Lect. Notes Comput. Sci, vol. 3896 LNCS, no. March, pp. 811–828, 2006,
44. B. Yu, G. Li, K. Sollins, and A. K. H. Tung, “Effective keyword-based selection of relational databases,” Proc. ACM SIGMOD Int. Conf. Manag. Data, no. July 2015, pp. 139–150, 2007,
45. G. A. Mihaila, L. Raschid, M.-E. Vidal, and M.-E. Vidal, “Using Quality of Data Metadata for Source Selection and Ranking.,” 2000.