Inhalt | (f) -> fett gedruckt in Originalmap

dsci-mindmap-ws2023

Barton, T., Müller, C. (2021). Data Science: Vom Begriff zur Anwendung. In: Barton, T., Müller, C. (eds) Data Science anwenden. Angewandte Wirtschaftsinformatik. Springer Vieweg, Wiesbaden. https://bibaccess.fh-landshut.de:2188/10.1007/978-3-658-33813-8_1

kw42_1.1 Was ist Data Science?

kw42_1.2 Was ist und was macht ein Data Scientist?

- Tab. 1.1 Themenbereiche für Qualifikationen von Data Scientists

Spezialisierungen:

- Data-Business-Person: Eine Person mit Fokus auf Qualifikationen für Business
- Data Creative: Eine Person, bei der die Qualifikationen ungefähr gleichmäßig auf die fünf Themenbereiche verteilt ist
- Data Developer: Eine Person mit Fokus auf Qualifikationen für Programmierung
- Data Researcher: Eine Person mit vertieften Qualifikationen für Statistik

Gute Data Scientists sollen

- über technische Expertise verfügen
- neugierig sein
- Problem in Hypothesen aufschlüsseln
- Storytelling betreiben
- Probleme kreativ und und auf unterschiedliche Weise anzugehen

KW 42_1.3 Einführung in Data Science:

- hier nicht wichtig

kw42_1.4 Systeme, Werkzeuge und Methoden

- hier nicht wichtig

Data-Science Anwendungsbereiche

Integration erneuerbarer Energien

- Energiewende mit dem Ausstieg aus der Kernenergie und damit verbundenen Herausforderungen

Machine Learning für die Energiemanagementoptimierung

- Optimierung einer Klimatisierungsanlage mithilfe von Data Science

Text Mining bei einer wissenschaftlichen Literaturauswertung

- Extraktion von Schlüsselwörtern zur Beschreibung von Inhalten

Identifikation relevanter Zusammenhänge in Daten mit maschinellem Lernen

- Zusammenhang zwischen Konfigurationen von Produkten/Infrastruktur und Fehlern

Data Science Was ist das ?

Dt. Datenwissenschaften

Das Filtern von Daten, um bestimmte Prozesse zu optimieren oder automatisieren

- - Datenanalyse ab den 1960er Jahren
- - Verbreitung in Unternehmen in den 1990er Jahren
- - Zunehmendes Datenvolumen durch Digitalisierung

Der Bereich der Datenwissenschaft befasst sich mit:

- - Der Analyse von (großen) Datenmengen - Der Identifizierung von Anomalien in den Daten - Der Vorhersage von zukünftigen Ereignissen
    - - Der Analyse von (großen) Datenmengen
    - - Der Identifizierung von Anomalien in den Daten
    - - Der Vorhersage von zukünftigen Ereignissen

Statistik + Informatik = Data Science

Kernbereiche von Data Science

- Data Engineering
- Data Analytics
- Data Prediction
- Maschinelles Lernen

Die neue Definition von Data Science

- basierend auf einen interdisziplinären Ansatz aus dem Jahr 2017 Data Science = (Statistik + angewandte Informatik + Computing + Kommunikation + Soziologie + Management | (Daten + Umgebung + Denkweise))

Ethik und Data Science

- Die Beurteilung sozialer Aspekte basiert auf moralische Prinzipien

Was ist und was mach ein Data Scientist

Spezialisierungen

- Data-Buisness-Person
    - Eine Person mit Fokus auf Qualifikation für Buisness
- Data Creative
    - Eine Person, bei der die Qualifikationen ungefähr gleichmäßig auf die fünf Themenbereiche verteilt ist
- Data Developer
    - Eine Person mit Fokus auf Qualifikationen für Programmierung
- Data Researcher
    - Eine Perosn mit vertieften Qualifikationen für Statistik

Was sollten gute Data Scientists können?

- über technische Expertise verfügen, die beispielsweise über ein naturwissenschaftliches Studium nachgewiesen werden kann
- neugierig sein mit einem Verlangen, zu entdecken und in die Tiefe zu gehen, um ein Problem in Hypothesen aufzuschlüsseln, die getestet werden können
- **Storytelling** betreiben, indem sie Daten dazu verwenden, um eine Geschichte zu erzählen und diese effektiv zu kommunizieren

Welche Qualifikationen sind dafür benötigt?

- Business/Produktentwicktlung
    - - Buisness
    - - Produktentwicklung
- Machine Learning/Big Data
    - - Big Data und verteilte Daten
    - - Machine Learning
    - - Strukturierte Daten

    - - Unstrukturierte Daten
- Mathematik/Operation Research
    - - Algorithmen
    - - Bayes'sche Statistik und Monte-Carlo-Methoden
    - - Grafische Modelle
    - - Mathematik
    - - Optimierung
    - - Simulation
- Programmierung/Systemadministration
    - - Back-End-Programmierung
    - - Front-End-Programmierung
    - - Systemadministration
- Statistik und Visualisierung
    - - Statistik
    - - Umfragen und Marketing
    - - Visualisierung

Data Science: Vom Begriff zur Anwendung

1.3 Einführung in Data Science

- Einführung in Data Science in Kapitel 2
- Ethische Betrachtungen sind eine immer größere Rolle in der digitalen Transformation von Unternehmen
- digitale Transformation führt zur Implementierung technologischer Lösungen zur Unterstützung der Entscheidungsfindung
- Untersuchungen zum Scheitern von Data-Science-Projekten in Kapitel 5

1.4 Systeme, Werkzeuge und Methoden

- "Empfehlungssysteme und der Einsatz maschineller Lernverfahren" von A. Peuker und T. Barton
    - Grundlagen und Einsatz von Empfehlungssysteme
- vergleich BI-Systeme und die Funktionalität aus dem Bereich machine Learning für Fachanwendungen

1.1 Was ist Data Science?

- Schnittmenge dreier Mengen, jede eine Kompetenz von Data Scientists
    - Hacking-Fähigkeiten
    - mathematische, statische Kompetenzen

        - substanzielle Kompetenzen
- vier Kernbereiche für die acatech
    - Data Engineering
    - Data Analytics
    - Data Prediction
    - maschinelles Lernen
- neuere Definition basierend auf interdisziplinärem Ansatz
    - Data Science=(Statistik+angewandte Informatik+Computing+Kommunikation+Soziologie+Management | (Daten+Umgebung+Denkweise)
- Data Science stützt sich auf
    - angewandte Informatik
    - Computing
    - Kommunikation
    - Management
    - Soziologie (soziale Aspekte)
- Moral
    - Beurteilung von sozialen Aspekten
    - Gesamtheit feststellbarer Verhaltensweisen, Verhaltensnormen und verhaltensbezogener Eistellungen und Werturteile
    - Gegenstand der Ethik

1.2 Was ist und was macht ein Data Scientist?

- Attraktivster Job des 21. Jahrhunderts
- 2015 Chief Data Scientist ernannt
- Was macht er und welche Qualifikationen braucht er?
    - Business und Produktentwicklung
    - Machine Learning/Big Data
        - Big Data und verteilte Algorithmen
        - Machine Learning
        - Strukturierte Daten
        - Unstrukturierte Daten
    - Mathematik/operation Research
        - Algorithmen
        - Bayes'sche Statistik/Monte-Carlo-Methoden
        - Grafische Modelle
        - Mathematik
        - Optimierung
        - Simulation

        - Programmierung/Systemadministration
            - Back-End-Programmierung
            - Front-End-Programmierung
            - Systemadministration
        - Satistik/Visualisierung
            - Statistik
            - Umfragen und Marketing
            - Visualisierung
- folgende Spezialisierungen:
    - Data-Business-Person: Fokus Business
    - Data Creative: Fokus gleichmäßig aufgeteilt
    - Data Developer: Fokus Programmierung
    - Data Researcher: vertieft auf Statistik

1.5 Anwendungen

- erneuerbare Energien (Kap. 9)
- Optimierung des Energiemanagements (Kap. 10)
- wissenschaftliche Literaturauswertungen (Kap. 11)
- zusammenhänge in Daten mit maschinellem Lernen identifizieren (Kap. 12)
- Kundenzufriedenheit in der Automobilindustrie und Fahrerassystenzsystementwicklung (Kap. 13)

http://jbusse.de/dsci-101/dsci-101-quellen.html#term-Zweig-2018

Katharina A. Zweig: Wo Maschinen irren können. Verantwortlichkeiten und Fehlerquellen in Prozessen algorithmischer Entscheidungsfindung. Bertelsmann Stiftung (Hrsg.), 05.02.2018, DOI: 10.11586/2018006, https://www.bertelsmann-stiftung.de/de/publikationen/publikation/did/wo-maschinen-irren-koennen > Download (pdf)

Zweig-2018-Abb-4

Abbildung 4 aus Zweig 2018, S. 21, als Mindmap ... hier ein Versuch, die Abb. als KnowledgeGraph in einer Mindmap darzustellen

Phase 1: Algorithmendesign und Implementierung

- REL_hatFehler
    - handwerkliche Fehler
        - Je mehr Anwender es gibt, desto wahrscheinlicher ist es, dass ein Fehler entdeckt wird

        - Um Fehler erkennen zu können, ist es vor allen Dingen wichtig zu wissen, wie der Algorithmus in welchem Fall reagieren sollte – die Problemspezifikation muss also bekannt sein.
        - Je mehr Personen Zugang zum Code haben, desto wahrscheinlicher ist es, dass einem von ihnen ein Fehler auffällt.
- REL_hatAkteur
    - Wissenschaftler/Informatiker

Phase 2: Methodenauswahl

- BT
    - Operationalisierung
        - REL_hatFehler
            - b Fehlende Passung von Operationalisierung und Daten
            - i Unpassende Daten für Fragestellung
            - e Mangelnde Datenqualität
    - Datensammlung
        - REL_hatAkteur
            - Datensammler (staatlich, wirtschaftlich, wissenschaftlich, NGOs)
    - Datenauswahl
        - REL_hatAkteur
            - Data Scientist
- REL_hatFehler
    - unpassende Methode

Phase 3: Konstruktion des Entscheidungssystems

- REL_hatFehler
    - j Unpassende Kombination von implementiertem Algorithmus und Daten
    - k Zu wenige Datenpunkte für Musteridentifikation
    - f Qualitätsmaß unpassend für Problemstellung

Phase 4: Einbetten in den gesellschaftlichen Kontext

- REL_hatFehler
    - c Fehlinterpretationen
    - a Fehlende Erklärbarkeit
    - h Unintendierte Nebenwirkungen durch Interaktion von System und Mensch

Phase 5: Re-Evaluierung des Entscheidungssystems

- REL_hatFehler
    - g Selbstverstärkende Feedbackschleifen

Phase 2-5

- REL_hatAkteur
    - Data Scientist

Phasen 3-5

- REL_hatAkteur
    - Entscheider (staatlich, wirtschaftlich, wissenschaftlich, NGOs)

Frick-DataGovernance

Messen und Beobachten

- Wie?
    - Kontinuierlich / Regelmäßig
    - Durch Zielsetzung + Aktueller Stand
- Wraum?
    - Verbesserung
    - Weiterentwicklung
    - Abweichungserkennung
    - Problemerkennung
    - Strategiereflexion

Technologie

- Was muss betrachtet werden?
    - Datenschutz
    - Datensicherheit
    - Datenqualitätsmanagement
- Wie gelingt die Umsetzung?
    - Bereitstellung geeigneter Werkzeuge
    - Schulungen für Mitarbeiter
    - Übersicht der Daten
    - Management der Metadaten
        - Bedeutung der Informationsobjekte
        - Prozessinformationen bzgl. Veränderung, Verknüprung, Zuordnung

        - Strukturangaben bzgl. Datentyp, Wertebereich, Qualität
        - Administrative Informationen über Erstellungszeitpunkt, Zugriffhäufigkeit, Berechtigung
- Richtiges Data-Management
    - Warum?
        - Zentrale Datenspeicherung für entwickler
        - Klassifizierung u. Anreicherung der Daten
        - DataLake erstellung durch unstrukturierte Daten
    - Data Lineage (Herkunft)
        - Aus aggregierten Datensätzen die Ursprünglichen Datensätze bestimmen
    - Data Catalog
        - Beschreibung der gespeicherten Daten

Kommunikation

- Wie?
    - Strukturierter Informationsaustausch
    - Frühzeitige Informationsweitergabe
    - Zielgruppenorientiert
    - Einhaltung von Richtlienien und Regeln
- Umsetzung
    - Kommunikationsplan
        - Wer benötigt Informationen
        - Wer ist verantwortlich
    - Trainingsplan
        - Schulungsplan für involvierte Gruppen/Personen

DataQuality Management

- Definition
    - Themenbereich der sich mit dem Arbeiten, mit qualitativ hochwertigen Daten befasst.
- Warum?
    - Daten haben wirtschaftliches Potenzial
    - Daten sind oft fehlerhaft, widersprüchlich, unvollständig oder veraltet
- Wie?
    - Validierung
    - Standardisierung
    - Bereinigung
    - Anreicherung

Data-Science / Wo Fehler passieren können

Wichtig:

- Fehler der Phase 4 (Einbettung in den gesellschaftlichen Kontext)
    -
        - Fehlinterpretationen, keine Erklärbarkeit, ausnutzen des Algorithmus
-
- Fehler der Phase 1 (Algorithmendesign und Implementierung)
    -
        -
            - Handwerkliche Fehler Treten im Design und der Implementierung auf
-
- Fehler der Phase 2 (Methodenauswahl)
    -
        - Operationalisierungsfehler Datenerhebung Qualität der Daten Veraltete Daten Methodenauswahl
-
- Fehler der Phase 3 (Konstruktion des Entscheidungssystems)
    -
        - Auswahl eines Qualitätsmaßes: Sensitivität, Spezifizität, Akkuratheit
-
- Fehler der Phase 5 (Re-Evaluation des Entscheidungssystems)
    -
        - System verstärkt -> mehr Feedback -> System verstärkt Selbstverstärkende Feedbackschleife

Katharina Zweig (2018) "Wo Maschinen irren können"

Gefahren bei Entscheidungssystemen? Weapons of "Math Destruction"?

- Intransparenz
- Skalierbarkeit
- Schadenspotential

Lösungsvorschläge

- Algorithmen-TÜV
- Data Science Berufsethik

- Beipackzettel für Algorithmen
- Validierung und externe Beforschbarkeit

Fazit

- Komplexität und Fehleranfälligkeit
- Encoding Values?

http://jbusse.de/dsci-101/dsci-101-quellen.html#term-Frick-2021b

Frick, D. (2021). Data Governance. In: Frick, D., et al. Data Science. Springer Vieweg, Wiesbaden. https://bibaccess.fh-landshut.de:2188/10.1007/978-3-658-33403-1_6

Was ist Data Governance?

Assets

Roles, Tasks & Responsibilities

Processes

Architecture & Tools

Security

Compliance

6.1.2 Datenstrategie

Perspektiven

Treiber für Data Governance

6.2 Data Governance Framework

6.2.1 Strategie

6.2.2 Aufbauorganisation

- Rollen im Data Governance

6.2.3 Richtlinien, Prozesse und Standards

- Wirkung auf die verschiedenen beteiligten Elemente (Systeme, Menschen, Prozesse, Daten)

6.2.4 Messen und Beobachten

6.2.5 Technologie

- **Erklären Sie:**
    - Metadaten
    - Data-Lineage, Data Provenance
    - ETL
    - Taxonomie
    - Data Catalog
    - Data Lake

Metadaten: Metadaten sind Daten, die Informationen über andere Daten bzw. Dateien bereitstellen (https://denisenoetzel.de/was-sind-metadaten/#:~:text=Metadaten%20Beispiele&text=Kamerafoto%20>>%20Dateiname%2C%20Dateigröße%2C,Dateiname%2C%20Title%20Tag%2C%20ALT%20Tag)
Data-Lineage: der Prozess des Verstehens, Aufzeichnens und Visualisierens von Daten auf dem Weg von Datenquelle zu Verbraucher (https://www.imperva.com/learn/data-security/data-lineage/#:~:text=Data%20lineage%20is%20the%20process,Data%20lineage%20process)
Data Provenance (dt. Datenherkunft): Bestimmung der ursprünglichen Datensätze, aus denen die Daten entstanden sind, in einem Data-Warehouse (https://de.wikipedia.org/wiki/Data-Lineage#:~:text=Datenherkunft%20(auch%20Data%20Provenance%20oder,aus%20denen%20sie%20entstanden%20sind.)
ETL: (steht für extract, transform, load)Datenintegrierungsprozess, der Daten von mehreren Quellen kombiniert und in ein einzelnes Data-Warehouse einlagert (https://www.ibm.com/topics/etl#:~:text=ETL%2C%20which%20stands%20for%20extract,warehouse%20or%20other%20target%20system.)
Taxonomie: Weg, Daten zu organisieren und zu klassifizieren. Gruppiert nach Charakteristiken, Attributen und Beziehungen zueinander (https://amplitude.com/explore/data/what-data-taxonomy)
Data Catalog: digitales Inventar bzw Verzeichnis, das als „single source of trust“ sämtliche Unternehmensdaten enthält. Ziel: Qualität und Geschwindigkeit der Datennutzung erhöhen (https://www.talend.com/de/resources/what-is-data-catalog/#:~:text=GBei%20einem%20Data%20Catalog%20(dt,Geschwindigkeit%20der%20Datennutzung%20zu%20erhöhen.)
Data Lake: zentrales Repository zum Speichern, Verarbeiten und Sichern großer Mengen an Daten, egal ob strukturiert oder nicht (https://cloud.google.com/learn/what-is-a-data-

lake?hl=de#:~:text=Ein%20Data%20Lake%20ist%20ein,strukturierter%2C%20semistrukturierter%20oder%20unstrukturierter%20Daten.)

6.3 Data Quality Management (DQM)

Prozessbereiche

Data-Profiling-Analyse

Data-Governance-KW44

Definition

- Rahmenwerk für Umgang mit Daten im Unternehmen

Inhalt

- Richtlinien
    - für Schutz
    - für Sicherheit
    - für Qualität
- Einhaltung der rechtlichen Vorgaben

Dimensionen

- Assets
    - Definition von Daten
- Roles, Tasks, Responsibilities
    - Festlegung der Rollen
        - Data Owner
            - Senior-Manager mit Wissen über Datensemantik
        - Data Steward
            - Mitarbeiter aus Fachgebiet Daten und IT
        - Data Custodian
            - Spezialisiert auf Vermeidung von Datenverlust/-verfälschungen
    -
- Processes
    - Überwachung der internen Prozesse und Übertragungen
- Architecture
    - setzt Standards für technische Umsetzung
- Security

        - Standards von Datensicherung
        - Zugriffsrechte
        - bestimmen Vorgehensweise bei Sicherheitsverstößen
- Compliance
        - Einhaltung des Datenschutzes

Perspektiven

- System
        - Regeln für Datenarchitektur
        - technische Komponente
- Prozess
        - Datenmanagement
                - Datenerhebung
                - Datenveräußerung
                - Datenlöschung
- Strategie
        - Fokus auf Optimierung
        - Fokus auf Entwicklung neuer datengetriebener Geschäftsmodelle

Werte

- Datennutzer
        - Flexibilität
        - Agilität
        - zeitnah
        - selbstständig Daten durchforsten
- Datenanbieter
        - Konsistenz
        - Transparenz
        - Verfügbarkeit

Bestandteile

- Vision
        - schwammig, ungefähres Ziel
- Mission
        - Rolle im Unternehmen
        - Regeln entwickeln
- Ziel
        - Klare, messbare Vision
- übergeordnetes Ziel

    - Maximierung des geschäftlichen Nutzens
    - Übereinstimmung mit Unternehmenszielen

Häufige Prozesse

- Aligning Policies, Requirements and Controls
- Establishing Decision Rights
- Establishing Accountability
- Performing Stewardship
- Managing Change
- Defining Data
- Resolving Issues
- Specifying Data Quality Requirements
- Providing Stakeholder Care
- Communications and Program Reporting
- Measuring and Reporting Value

Schritte zur Einführung

- Ermittlung des Status Quo im Datenmanagement
- Ziele definieren
- Ursprüngliches Konzept + Roadmap
- Zustimmung von Stakeholder & Sponsor
- Konzept ausarbeiten und transformieren
- Roadmap umsetzen
- Ausweitung in andere Bereiche
- Stabilisieren und Verbessern

NEU 2024-01-16: Jens Kaufmann, Kap. 11: Fundamentale Analyse- und Visualisierungstechniken

jeweils ganz kurz erklären können

- Boxplot
- 11.2 Lineare Regression
- 11.3.1 k-Nearest-Neighbors
- 11.3.2 Naive Bayes
- 11.3.3 Entscheidungsbäume
- 11.3.3 Entscheidungsbäume
- 11.4 Clustering-Verfahren
    - 11.4.1 Hierarchische Verfahren
        - Dendrogramm

    - 11.4.2 Partitionierende Verfahren
        - k-means
- 11.5 Assoziationsanalyse

- Boxplot: kurze visuelle Zusammenfassung der Variabilität von Werten in einem Dataset. Ausreißer können Fehler oder Ungewöhnliches in Daten aufdecken
- (https://doc.arcgis.com/de/insights/latest/create/box-plot.htm#:~:text=Boxplots%20stellen%20eine%20kurze%20visuelle,ungewöhnliche%20Vorkommnisse%20in%20Daten%20aufdecken.)
- Lineare Regression: Spezialfall der Regressionsanalyse; Versuch, eine beobachtbare, abhängige Variable durch unabhängige Variablen zu erklären (lineares Modell)
- (https://de.wikipedia.org/wiki/Lineare_Regression#:~:text=Die%20lineare%20Regression%20(kurz%3A%20LR,(kurz%3A%20LM)%20angenommen.)
- k-Nearest-Neighbors: Klassifikationsverfahren, bei dem eine Klassenzuordnung unter Berücksichtigung seiner nächsten k-Nachbarn vorgenommen wird. Der Teil des Lernens besteht aus simplem Abspeichern der Trainingsbeispiele, was auch als „lazy learning“ bezeichnet wird. Datennormalisierung kann Genauigkeit des Algorithmus erhöhen (https://de.wikipedia.org/wiki/Nächste-Nachbarn-Klassifikation)
- Naive Bayes: Technik des maschinellen Lernens für Klassifikationen. Objekte können in mehrere Klassen eingeteilt werden. Lernt durch Analyse von Daten, die richtig klassifiziert sind
- (https://www.alexanderthamm.com/de/data-science-glossar/naive-bayes/#:~:text=Naive%20Bayes%20ist%20ein%20probates,oder%20mehr%20Klassen%20eingeteilt%20werden.)
- Entscheidungsbäume: siehe ID3 aus KI
- Clustering-Verfahren: Verfahren zur Entdeckung von Ähnlichkeitsstrukturen und Datenbeständen. Die gefundenen Gruppen werden als Cluster bezeichnet.
- (https://de.wikipedia.org/wiki/Clusteranalyse)
- Hierarchische Verfahren bei Clusteranalyse:
- - divisive Verfahren (Top-Down-Verfahren): erst ein großer Cluster, immer weiter aufgeteilt in kleinere Cluster
- - agglomerative Verfahren (Bottom-Up-Verfahren): erst jedes Objekt eigener Cluster, immer weiter zusammengeführt zu größeren Clustern
- (https://de.wikipedia.org/wiki/Hierarchische_Clusteranalyse)
- Dendrogramm: siehe nächste Seite
- Dendrogramm-Beispiel:
-
- Partitionierende Verfahren: Zahl der Cluster muss festgelegt werden (Nachteil). Clusterzentren werden solange verschoben, bis sich Zuordnung der Beobachtungen zu Clusterzentren nicht mehr ändert. Vorgegebene Fehlerfunktion wird minimiert. Objekte können während Verschiebung der Clusterzentren Clusterzugehörigkeit wechseln (Vorteil)

(https://de.wikipedia.org/wiki/Clusteranalyse#:~:text=Partitionierende%20Clusterverfahren,-Partitionierende%20Clusteranalyse&text=Clusterzentren%20nicht%20mehr%20verändert%2C%20wobei,Clusterzentren%20ihre%20Clusterzugehörigkeit%20wechseln%20können.)

- k-means-Algorithmus: (Jens Kaufmann selbst)
- 1. jedem Datenpunkt einen (zufälligen) Cluster zu (Startlösung)
- 2. Berechne den Mittelpunkt aller Cluster
- 3. Weise jedem Datenpunkt dem Cluster zu, dessen Mittelpunkt er am nächsten liegt
- 4. Wenn die Lösung sich nicht mehr ändert, ist der Zielzustand erreicht, ansonsten gehe wieder zu Schritt 2
-
- Assoziationsanalyse: Suche nach starken Regeln. Assoziationsregeln beschreiben Korrelationen zwischen gemeinsam auftretenden Daten
- (https://de.wikipedia.org/wiki/Assoziationsanalyse)

KW 45, Thema: {term}□□□□□□∩2021

http://jbusse.de/dsci-101/dsci-101-quellen.html#term-Kaufmann-2021

Kaufmann, J. (2021). Fortgeschrittene Verfahren zur Analyse und Datenexploration, Advanced Analytics und Text Mining. In: Frick, D., et al. Data Science. Springer Vieweg, Wiesbaden. https://bibaccess.fh-landshut.de:2188/10.1007/978-3-658-33403-1_12

jeweils kurz erklären und einordnen könnnen

- Recherchiere: Unterschied überwachtes Lernen / unüberwachtes Lernen?
- 12.2 Datenexploration und -darstellung
- 12.3 Principal Component Analysis
- 12.4 Random Forests
- 12.5 Logistische Regression
    - Unterschied Regression und Logistische Regression?
- 12.6 Entscheidungsbewertung
    - Konfusionsmatrix
- 12.7 Zeitreihenanalyse
- 12.8 Text Mining
    - Bag of Words
    - Term Frequency – Inverse Document Frequency (TFIDF)
    - Kosinus-Ähnlichkeit

KW45_Vortrag-dsci-kaufmann_2023_11_06

(JBusse: habe hier die Formeln entfernt, nicht relevant für die Klausur)

1. Principal Component Analysis

- * Reduzierung komplexer Datenbestände
- * Kombination von Variablen
- * Erzeugung künstlicher neuer Variablen

2. Random Forest

- * Besteht aus mehreren Entscheidungsbäumen
- * Verbessert Klassifikationsgüte

3. Logische Regression

- * Ermöglicht Schätzung von Wahscheinlichkeiten
- * Basiert auf Umrechnung von Wahrscheinlichkeiten zu Chancen
- * Verwendet S-förmige Sigmoidfunktion

4. Entscheidungsbewertung

- * Erfolgt durch Analyse von Konfusionsmatrizen
- * Vergleich von vorhersagen und tatsächlichen Werten

5. Zeitreihenanalyse

- * Ermöglicht untersuchung von zeitlichen Entwicklungen
- * Schätzung zukünftiger Werte

6. Text Mining

- * Analyse natürlichsprachlicher Texte
- * Ähnlichkeit zwischen texten durch Kosinus-Ähnlichkeit

KW 45, Fortgeschrittene Verfahren zur Analyse und Datenexploration, Advanced Analytics und Text Mining

Hauptgruppen des Data Mining

- Klassifikation (Objekte zuordnen zu Klassen)
- Segmentierung/Clustering (Objekten in Gruppen einteilen)
- Prognose (auf Basis bekannter Werte)
- Assoziationsanalyse (Zusammenhang einzelner Elemente erkennen)

Datenexploration und -darstellung

- zielführende graphische Darstellung der Daten
    - Für Menschen ist dies angenehmer zu analysieren als Tabellen mit numerischen Werten
    - betrachte große Datenmengen explorativ, aber nicht planlos
    - Erkennung von Mustern auf höherer Ebene mithilfe zusammenfassender Darstellungen
    - Können damit Hypothesen erstellen und passende Analyseverfahren der Daten auswählen

Logistische Regression

- Die Zuordnung eines Datenpunktes zu einer Klasse wird ein Wahrscheinlichkeitswert gegeben
- Zieht eine Entscheidungsgrenze (Linie) durch Datenpunkte
    - Entfernung von Entscheidungsgrenze bestimmt die Wahrscheinlichkeit der korrekten Klasseneinordnung
    - Klassen einordnen auf beiden Seiten der Entscheidungsgrenze

Random Forest

- entsteht aus mehreren Entscheidungsbäumen zusammengefügt
    - Training Set (für eigentlichen Modell-Erstellung)
    - Validation Set (zur Verbesserung des Modells)
    - Test Set (zur Qualitätermittlung)
- hat höhere Qualität als einzelne Entscheidungsbäume

Zeitreihenanalyse

- beschreibt kausaler Zusammenhänge zwischen Zeitreihen
    - aus Längsschnittdaten der Variablen über einen Zeitraum
    - zerlege Längsschnittdaten in einzelne Komponenten
        - Trendkomponente, beschreibt langfristige Entwicklungen
        - saisonale Komponente, beschreibt wiederkehrende Muster
        - Zufallskomponente, ist Restgröße der Datenveränderung
        - Angabe Konfidenzintervalle, da mit zeitlicher Entfernung zum letzten Datensatz die Genauigkeit abnimmt
        - Visualisierung mithilfe Punkt-, Linien- oder Säulendiagramme
- ist eine Prognose zukünftiger Werte

Text Mining

- **bag-of-words-Ansatz zum Strukturieren eines Texts**

    - Für jedes analysierte Dokument hat die Tabelle eine Zeile
    - jedes Wort hat eine Spalte
        - In jeder Spalte wird die Häufigkeit des Worts notiert
    - diverse Fehlerquellen, welche Analyse erschweren, müssen entfernt werden
    - Ähnlichkeit von Dokumenten durch Kosinus des Winkels der entsprechenden Vektoren beschreiben

Entscheidungsbewertung

- prüft die Qualität von Modellen
    - Confusion Matrix
        - False Positive Fraction, also Fehleinschätzungen
        - True Positive Fraction, also korrekte Einschätzungen
        -
    - die Receiver- Operating-Characteristics-Kurve (ROC-Kurve) stellt Fractions in Abhängigkeit zu Schwellwert dar
        - je weiter ROC-Kurve von Diagonalen entfernt, desto präziser das Modell
        - größere Fläche unter ROC-Kurve bedeutet besseres Modell

Hauptkomponentenanalyse

- auch „Principal Component Analysis“ (PCA) genannt
- ermöglicht vielen Variablen in Graphen nachvollziehbarer darstellen
    - kombiniere bestehende Variablen zu einer neuen Variable
    - hat gleiche Eigenschaften in geringerer Dimension
    - ermöglicht Graphik-darzustellung im 3-dimensionalen-Raum von neuen Variablen und ihren Gruppen

Einleitung in Thema

- Fragen für Datenanalyseverfahren
    - welche Daten stehen zur Verfügung vor?
    - welche Fragestellung sollen beantwortet werden?
    - welche Methode der Datenanalyse Verfahren sind sinnvoll?
- Vorgehensweise in Datenanalyseverfahren
    - 4 Hauptgruppen des Data-Mining
    -
    - erster Ansatz Datenexploration und -darstellung
    - zweiter Schritt Hauptkomponentenanalyse
        -
        - danach weitere Datenanalyseverfahren

http://jbusse.de/dsci-101/dsci-101-quellen.html#term-Maierhofer-2021

Maierhofer, C.R. (2021). Information Data Models: Das Fundament einer guten Information Strategy. In: Frick, D., et al. Data Science. Springer Vieweg, Wiesbaden. https://bibaccess.fh-landshut.de:2188/10.1007/978-3-658-33403-1_9

- Was ist eine native Data Science Strategie?
- 9.1 Drei Thesen aus Sicht eines Praktikers
    - Chaos
    - Hierarchien
    - selbstbeschützend
    - growth midset, fixed mindset
    - Knechtschaft der Applikationen / autonome Existenz von Daten
- 9.3 Das Heute und seine Hürden
- 9.4 Wie es dazu gekommen ist
- 9.5 Die Enterprise Architektur
- 9.6 Drei Formen der Informations-Architektur und deren Auswirkungen
    - 9.6.1 Das Gestern und leider noch das Heute. Der anwendungszentrierte Ansatz (The Application Centric Approach)
    - 9.6.2 Das Heute und die Morgendämmerung, der datengesteuerte Ansatz (The Data Driven Approach)
    - 9.6.3 Das überfällige Übermorgen, die datenzentrische Architektur (The Data Centric Architecture)
        - Data-Centric Manifesto
        - Abb. 9.4 Data Centric Architecture

Information Data Models - Herausforderungen und Lösungen

Das Heute und seine Hürden

- - Aktuelle Probleme: Diskrepanz zwischen Selbstwahrnehmung und Realität in Unternehmens-Informationssystemen. - Interne Herausforderungen: Schwierigkeiten bei Leistungsverrechnung und Mangel an aussagekräftigen Performance Indicators.
- - Abteilungsübergreifende Probleme: Komplikationen bei der Erfassung und Verarbeitung von Informationen zwischen verschiedenen Bereichen.
- - Analyse: Notwendigkeit von mehr Transparenz und Flexibilität in den bestehenden Systemen.

Die Enterprise Architektur

- - Moderne Ansätze: Notwendigkeit der Anpassung an das aktuelle Geschäftsmodell und datenzentrische Architekturen.
- - Vorteile: Erhöhte Agilität und Anpassungsfähigkeit, effizientere Ressourcennutzung durch zentralisierte Datenstrukturen.

Abschluss und Fazit

- - Zusammenfassung: Wichtigkeit der Modernisierung der Informationsarchitektur in Unternehmen.
- - Ausblick: Bedeutung effektiver Data Science Strategien für den zukünftigen Erfolg von Unternehmen.

Wie es dazu gekommen ist

- - Historischer Kontext: Entwicklung der IT-Abteilungen von Basisservice-Anbietern zu strategischen Partnern.
- - Folgen: Budgets und Kontrolle verschoben sich in Richtung Fachbereiche, Applikationszentrierte Architekturen entstanden

Information Data Models

Informationen

Drei Thesen aus Sicht eines Praktikers

- Allgemein
  - Chaotische Informations-Architektur
    - Anerkennung des Problems
    - Willen zur Veränderung
  - Hirarchie der Organisationsstruktur
    - Beschränkung
      - Weiterentwicklung
      - automatische Verarbeitungsmöglichkeiten
- Bedeutung des Mindset
  - Growth Mindset
    - Wille zur Veränderung
    - Bringt Fortschritt
  - Fixed Mindset
    - Birgen in Bequemlichkeit
    - Fördert konservative Struktur

        - Native Data Science Strategie
            - Fundamentale Veränderung
            - Autonome Datenverarbeitung
            - Autonome Existenz von Daten
- **Informationen als entscheidender Wirtschaftsfaktor**
        - KI als Paradebeispiel
            - Datensammlung im autonomen Fahren
            - Datensammlung in LMMS
        - Unternehmen sollten Daten höchste Priorität einräumen

 # KW 47, Thema: {term}��h����2021

http://jbusse.de/dsci-101/dsci-101-quellen.html#term-Schmitz-2021

Schmitz, U. (2021). Big Data. In: Frick, D., et al. Data Science. Springer Vieweg, Wiesbaden. https://bibaccess.fh-landshut.de:2188/10.1007/978-3-658-33403-1_1

- 1.1 Grundlagen
        - Die 5 V
- 1.2 Architektur und Bausteine
        - Abb 1.1 Klassifizierung von Big Data-Technologien
            - AUFGABE: Klassifizierung rekonstruieren
- 1.3 Datengetriebene Geschäftsmodelle
- 1.4 Exemplarische Einsatzmöglichkeiten

Big Data

Architektur und Bausteine

- Baustein: Data Governance
        - Internationale rechtliche Rahmenbedingungen
            - EU-Datenschutzgrundverordnung
            - Internationale Unterschiede
                - EU vs US
                - Notwendigkeit Internationaler Standards
        - Berücksichtigung der Interessen einer Person
            - Deutsches Bundesdatenschutzgesetz
            - Wichtigkeit und Bedeutung personenbezogener Daten
        - Anonymisierung und Pseudonymisierung
            - Anonymisierung
            - Pseudonymisierung

    - AOL Skandal
- Baustein: Datenkonnektivität
    - ETL-Prozess

Datengetriebene Geschäftsmodelle

- Optimierung
    - Konzept: Bessere Auswertung existierender Datenbestände
    - Beispiele: Konvertierung alter Datenpools in neue Formate, etc.
- Monetarisierung
    - Konzept: Schaffen neuer Produkte mit bestehenden Daten
    - Beispiele: Analyse-Dienste basierend auf Suchverhalten, etc.
- Leverage
    - Konzept: Bestehende Geschäftsmodelle durch Daten verbessern
    - Beispiele: Intelligente Mautsysteme, etc.
- Disrupt
    - Konzept: Produkte durch Sammeln neuer Datenbestände
    - Beispiele: Facebook, etc

Big Data-Geschäftsmodelle

Analytics-as-a-Service

- Konzept: Bereitstellung von Analysen und Prognosen
- Beispiele: Wetter, Social Media, etc.

Data-as-a-Service

- Konzept: Sammeln, Aggregieren von Daten
- Beispiele: Online-Werbung, Cookies, etc.

Data-infused Products

- Konzept: Aufwertung bestehender Produkte durch Daten
- Beispiele: Intelligente Stromzähler, etc.

Datenmarktplätze und Daten-Aggregatoren

- Konzept: Plattformen für Verkauf und Nutzung von Daten
- Beispiele: Marktforschungs- und Beratungsunternehmen, etc.

1.4 Exemplarische Einsatzmöglichkeiten

Social Media

- Web 2.0

- Mitgestaltung von Inhalten durch Nutzer in sozialen Netzwerken, Blogs usw..
- Nutzung für Marketing und PR, interne Kommunikation im Unternehmen. z.B. über Twitter, Facebook oder interne Wikis.
- Entwicklung von Strategien zur Positionierung von Unternehmen auf Plattformen. Hauptsächliche Inhalte davon sind:
    - Bekanntmachung von Inhalten
    - Kontakt zu Nutzern

Proaktiver Ansatz

- Setzt auf direkte Kommunikation (zB. Facebook, Blogs).
- Kunden können aktiv in Marketingaktivitäten wie Produktgestaltung durch Crowdsourcing einbezogen werde.
- Dieser Ansatz zielt auf langfristige Kundenbeziehungen und verspricht größeren Erfolg im Vergleich zum reaktiven Ansatz.

Reaktiver Ansatz

- Grundsätzlich abwartende Haltung.
- Überwachung von Social Web nach Feedback zum Unternehmen.
- Unternehmen reagiert gezielt auf Kritik, entgegenwirken und aufklären.
- Hauptsächlich um Nutzverhalten zu beobachten und Feedbacks zu überwachen.
- Zudem werden Social Media Guidelines erstellt, um Mitarbeitern klare Richtlinien für ihr Verhalten im Social Web zu geben, einschließlich Kommunikation mit Dritten, Datenschutz und Urheberrecht.

Marketing und Vertrieb

- Unternehmensbeispiel Telefonica
    - Telefónica, ein spanischer Telekommunikationskonzern führte Smart Steps ein:
        - sammelt ortsbezogene Daten von Nutzern
        - anonymisiert die Daten
        - verkauft an Dritte
    - Nutzung: Besucherzahlen zu bestimmten Zeiten um Personal zu optimieren. Mobiltelefonhersteller verbesserte dadurch Empfangsleistung.

Forschung und Entwicklung

- Unternehmensbeispiel UPS
    - Entwickelten Strategie zur Überwachung von Lieferungen, Routenoptimierung und Kostenreduzierung.

- Sensorsystem in jedem Fahrzeug:
    - Geschwindigkeit
    - Richtung
    - Benzinverbrauch
    - weitere technische Parameter
- kombiniert mit GPS-Daten ermöglicht dies:
    - Analyse von Fahrverhalten
    - Routenoptimierung
    - vorausschauende Wartung
- Folgen daraus:
    - Was zu einer Einsparung von 85 Millionen Meilen Wegstrecke pro Tag geführt hat, was etwa 30 Millionen Dollar pro Tag entspricht. Kunden haben zusätzlich mehr Einsicht in ihre eigenen Lieferdaten (z.B. Lieferzeit, Standort) was zu mehr Kundenzufriedenheit führte.

Finanz- und Risikocontrolling

- Unternehmensbeispiel United Overseas Bank
    - Prozess zur Bewertung des Gesamtrisikos wurde drastisch verbessert. Dabei werden über 100.000 marktrelevante Parameter analysiert.
        - Früher: 18 Stunden für ca. 8,8 Milliarden Berechnungen
        - Heute: wenige Minuten
    - Für die Problemlösung hat die Bank eine analytische Software-Lösung sowie eine In-Memory- Technologie eingeführt.
    - Diese Big Data-Technologien ermöglichen es sogar, neue marktrelevante Faktoren während der Berechnungen einzubeziehen.

Produktion, Service und Support

- Unternehmensbeispiel Vestas
    - Analyse für potentielle Standorte von Windkraftanlagen wurde drastisch beschleunigt und Stromerzeugungskosten pro kWh. wurden verringert. Weiterhin wurden die Ausfallzeiten der Anlagen durch die Berücksichtigung von materialbelastenden Turbulenzen minimiert.
        - Früher: mehrere Wochen
        - Heute: wenige Stunden
    - Die Analysen umfassen verschiedene Faktoren wie Geländehöhe, Satellitenbilder, Bewaldung, Stromnetzanbindung und historische Wetterdaten.

KW47 Big Data: Bausteine

Datenhaltung

Hadoop

- Open-Source Framework
- ermöglicht parallele Verarbeitung großer Datenmengen
    - mittles Map-Reduce-Methode
        - ermöglicht das Aufteilen großer Datenmengen in kleinere Teilmengen
- nutzt performante Computercluster
    - Netz aus miteinander verbundenen Computern mit einem Access-Point
    - Rechenlast für eine Aufgabe wird auf mehrere Computer verteilt
- keine festgelegte Struktur und Semantik der Dateien nötig
- Hadoop Distributed File System (=HDFS)
    - bringt Hochverfügbarkeit mit sich
        - auch bei Ausfall einzelner Komponenten bleiben alle Funktionen bestehen
        - die Daten werden dafür in dem Cluster gespeichert
- Vorteile
    - hohe und einfache Skalierbarkeit
    - Open-Source Framework
        - allgemeine Kosten sind niedriger als bei Software-Herstellern

Datenverarbeitung

- schnelle Verarbeitung mit In-Memory-Technologie
    - Daten werden nicht mehr auf der Festplatte gespeichert sondern im Arbeitsspeicher
        - ABER: im RAM speichern ist sehr aufwendig
        - LÖSUNG: Temperatur-Modell
            - Hot-Daten: oft verwendete Daten
                - speichern im RAM
            - Cold-Daten: selten verwendete Daten
                - speichern auf Festplatten

Datenverarbeitungsmethoden

Text-Mining

- Analyse von Fließtext(=unstrukturierte Datenmengen), um Muster zu erkennen
    - PROBLEM: jede natürliche Sprache hat eine andere Grammatik und Semantik
    - LÖSUNG: Natural Language Processing; führt die semantische Analyse des Fließtextes durch
    - Verwendungszweck: Social Media Marketing

Data Mining

- versch. Methoden, um Informationen aus den Daten zu erhalten
    - Segmentierung: Bildung von kleinen Gruppen
    - Abweichungsanalyse: Soll-Werte werden mit Ist-Werten verglichen und dementsprechend sortiert
    - Klassifikation: Daten in versch. Klassen aufteilen und sortieren
    - Prognose: Vorhersage auf bereits gewonnen Daten treffen
    - Assoziationsanalyse: Suche nach anwendbaren Regeln
    - Sequenzanalyse: Suche nach Relationen untereinander

Datenvisualisierung

- mit der Big-Data Ära brauchte man neue Darstellungstypen, um die Daten...
    - ...anschaulich zu gestalten
    - ...eventuelle Messfehler zu erkennen
    - Beispiele
        - Donut-Cloud
            - https://bibaccess.fh-landshut.de:2673/chapter/10.1007/978-3-658-33403-1_1#Fig4
        - Flare-Chart
            - https://bibaccess.fh-landshut.de:2673/chapter/10.1007/978-3-658-33403-1_1#Fig5
        - Dashboard

- https://bibaccess.fh-landshut.de:2673/chapter/10.1007/978-3-658-33403-1_1#Fig6

Textquelle

- https://bibaccess.fh-landshut.de:2673/chapter/10.1007/978-3-658-33403-1_1

KW 48, Thema

http://jbusse.de/dsci-101/dsci-101-quellen.html#term-Quix-2021

Quix, C. (2021). Data Engineering. In: Frick, D., et al. Data Science. Springer Vieweg, Wiesbaden. https://bibaccess.fh-landshut.de:2188/10.1007/978-3-658-33403-1_5

5.1 Aufgaben des Data Engineering

- CRISP-DM
- Data Wrangling

5.2 Architekturen zum Daten-Management

- Data-Warehouse
- ETL
- Data-Lake

5.3 Datenmodellierung und Metadaten-Management

5.4 Datenaufbereitung und Datenintegration

- Exploration & Profiling
- Harmonisierung & Bereinigung
- Identifikation der Quell-Schemata
- Abgleich der Quell-Schemata
    - Schema-Matching
- Definition eines integrierten Schemas
- Mapping der Quell-Schemata auf integriertes Schema
- Daten zusammenführen
- Feature-Selektion und -Konstruktion

5.5 Datenbank-Management-Systeme: SQL, NoSQL und Big Data

- NoSQL
    - Key-Value

- Dokument-orientiert
- Wide Column
- Graph-orientiert

5.4 Datenaufbereitung und Datenintegration

Datenintegration

Abgleichen der Datenschemata

- Ähnlichkeiten (Wissen, Korrespondenz) mit Shema Matching Tools erfassen
- Beziehungen
  - Vergleich von Zeichenketten
  - mit Wörterbuch vergleichen
- Datentypen
  - Ausschluss von Korrespondenzen
- Wertebreich
  - Histogramme analysieren
  - 2 Attribute = gleiche Werte (zB Alter)
- Struktur
  - Graph- oder Baumstruktur
  - Aus Ähnlichkeiten Nachbarn ableiten (Adresse = adresse)
- Referenzmodell
  - Ableiten durch Lokig oder Maschinen Learning
  - Deep Learning: gut für komplexes
  - Shema Matching: fehlende Trainingsdaten
- verschieden Ansätze = menschen müssen aber überprüfen

Integrierte Schemata

- Quellenorientiert
  - Vereinigungsmenge der Quellshemata
  - Berücksichtigung vorheriger Schritte
  - Unterstützung der Werkzeuge möglich
- Anwendungsorientiert
  - Ähnlich der Top-Down_Datemodellierung
  - Definiert durch geplante Anwendung
  - Vorteil: besser passende Datenmodelle, Informationslücken erkennbar

Mapping integrierter Schemata

- Daten aus Datenquellen extrahieren und einheitlich übernehmen
  - Definieren als Anfrage
  - über ein Werkzeug möglich

- Notwendige Vorarbeit schon getroffen
- Datenintegrationswerkzeuge
    - Datentransformation und Zusammenführung
    - Unterstützung bei einem Prozess aller Schritte

Datenzusammenführung

- Vorherige Schritte auf der Schemaebene
- Konkrete Zusammenführung von Datensätzen
- Record Linkage: welche Datensätze entsprechen dem selben Objekt

Data-Lake-Architecture

Definition

- Data-Lake-Architecture is a framework or approach to designing a central repository to store and manage data in its original format, without any predefined schema.
    - A database schema refers to the logical and visual configuration of the entire relational database.

However, that stability like in case with Data-Warehouse is not always the case with big data projects. Most big data systems rely on schema-on- Read concept in the foreground.In contrast, the Data-Warehouse system follows the schema-on-write approach.

- Schema-on-Read means that the data is initially stored without a predetermined schema.
- Schema-on-Write is a traditional approach where data is first structured and transformed before being loaded into a data storage system.
    - The schemas of the data sources and the data warehouse database are known before data is written to the Data-Warehouse database using ETL processes.
- However, the schema-on-write model is not suitable for big data because there is a larger number, more heterogeneity and greater agility in data sources
    - In contrast to Data-Warehouse systems, with data lake systems the data is transferred to the storage level of the system in its original form. Such an approach suits big data and NoSQL systems, which typically do not require the definition of a schema before data can be stored.
- Therefore, a different architecture should be chosen for data provision in big data projects that allows greater flexibility.
    - The data should be stored in the data lake in its original form and a transfer to a uniform scheme as with Data-Warehouse systems is not intended here.

Although in addition to the actual data, metadata should also be extracted from the data sources or recorded separately.

- Metadata is also important for query processing in the data lake system. An integrated query interface doesn't help if you don't know which data management systems contain the desired data.
- Metadata management is even more important in data lake systems than in data warehouse systems. While in data warehouse systems, the mostly relational database management systems can provide sufficient self-information about the schemas of their databases, this is not always the case in the context of Data-Lake-Systems due to the unstructured data.

Datenmodellierung und Metadaten-Management

Datenmodellierung

Top-Down-Ansatz

- konzeptuelles Datenmodell
- Datenmodell wird verfeinert
- Umsetzung als physisches Modell in einem Datenbank-System

Bottom-Up-Vorgehen

- existierende Datensätze
- ableiten von logischen Datenmodellen
- Beschreibung von semantischen Zusammenhängen in einem konzeptuellen Datenmodell

Data Profiling

- Schemaextraktion
    - es reichen Angaben die, beispielsweise für die Erstellung eines relationalen Schemas erforderlich sind
    - Erkennung von Attributen
        - Integer
        - string
        - usw.
- Data Profiling
    - Erkennung genauerer Wertebereiche oder Muster in Datensätzen
    - z.B. eine Spalte "Alter" hat nur Integer-Werte von 0 bis 120 oder eine Spalte "Datum" hat eine Zeichenkette mit dem Muster "DD.MM.YYYY"
- mit Data Profiling werden sehr schnell fehlerhafte Daten und Ausreißer erkannt => deswegen relevant für Datenaufbereitung

Datenmodell

- Erstellung eines logischen Datenmodells
    - direkte Auslesung von Schemata aus relationalen Datenbank-Systemen
- Beschreibung von Datensätzen
    - zumindest Strukturen, Verknüpfungen und Regeln bzw. Einschränkungen(Constraints) von Daten
- Modellierungssprache
    - J.Busse: heute überholt:
        - Data Definition Language(DDL)
        - SQL
        - XML Schema
    - **Unified Modeling Language(UML)**
    - J.Busse: **RDF(S)**
        - https://de.wikipedia.org/wiki/Resource_Description_Framework

Metadaten-Management

- Übersicht der vorhandenen Daten
- Aufbau und Verwaltung von Verzeichnis innnerhalb eines Unternehmens => Teil der Data Governance
- So ein Verzeichnis im Data-Lake-System Metadaten die dabei für einen Datensatz erfasst werden sollen:
    - Inhalt: Schlagworte, Themen, Beschreibung
    - Herkunft: Quellsystem, Kontext der Datenerfassung (z. B. Ort, Zeit)
    - Datenqualität: Messwerte für Datenqualitätseigenschaften (z. B. Vollständigkeit)
    - Kontakt: Ansprechpartner für die Datenquelle, Verantwortliche
    - Verfügbarkeit: Zugriffsmöglichkeiten, Lizenzinformationen, Nutzungseinschränkungen
- Metadatenmodell = individuell für jedes Unternehmen
    - Anpassung für individuellen Bedürfnisse mit bereits entwickelte, ausgereifte Modellen

Erste-Hälfte-Architekturen_zum_Daten-Management

Data-Management-Systeme in Unternehmen

- Heutzutage Einsatz vieler unterschielicher Systeme
- auch NoSQL -Datenbank-Management-Systeme

- Überwiegend jedoch relationale Systeme (SQL)
- Soll Ko-Existenz von verschiedenen Architekturen und Systemen gewährleisten

Data-Warehouse-systeme

- Sollen innerhalb eines Unternehmens die Daten aus verschiedenen operativen Systemen zusammenzuführen.
- Daten werden Durch ETL-Prozesse extrahiert und im Data-Warehouse bereitgestellt.
- Falls man Daten verschiedener Nutzergruppen trennen möchte, wird das Data-Warehouse in “ Data Marts ” unterteilt.
- Sind vor allem für Anwendungsfälle geeignet:
    - in denen die Datenstrukturen der Datenquellen und benötigten ETL-Prozesse sehr stabil sind und sich nur selten ändern
    - wenn man langfristige Zahlen miteinander vergleichen will

OLTP (On-Line Transaction Processing)

- z.B.Erfassung einer Bestellung oder Buchung einer Rechnung

OLAP (On-Line Analytical Processing)

- z.B. Analytische Anfragen, alle Verkaufsaktivitäten in einem Quartal

ETL-Prozesse (Extraktion-Transformation-Laden)

- Sorgen für Aufbereitung der Daten aus den heterogenen OLTP-Systemen
- Können heutzutage die Daten im Data-Warehouse in nahezu Echtzeit aktualisieren
- Erstellung sehr aufwendig

Datenbank-Management-Systeme: SQL, NoSQL und Big Data

Datenbank-Management-Systeme (DBMS)

Eigenschaften

- Über Schnittstelle mit definierter Sprache (z.B. SQL) Daten
    - Anlegen
    - Löschen
    - Ändern
    - Abfragen
- Unterstützung von Transaktionen
- Gewährleistung der Persistenz von Daten

Geschichte

- 1970er: Ursprung (RDBMS)
    - SQL als Datenbanksprache

        - Vorteile
            - Datenintegrität
            - Konsistenz
            - Transaktionssicherheit im Mehrbenutzerbetrieb
        - Probleme relationaler DBMS im Kontext verteilter Anwendugen
            - CAP-Theorem (Consistency, Availability, Partition Tolerance)
                - Anforderungen an verteiltes System
                - Gleichzeitige Gewährleistung aller drei Eigenschaften unmöglich → Priorisierung zweier Punkte
                - Partition Tolerance bei stark verteilten Anwendungen besonders wichtig (Fehlertoleranz gegenüber Netzwerkunterbrechungen!)
                - Problem: klassische relationale Systeme mit keiner oder nur eingeschränkter Verteilung fokussieren sich auf Consistency und Availability
            - Inkompabilität der Datenmodelle
                - Normalisierte Relationen (RDBMS)
                    - Normalisierung: Aufteilung großer, redundanter Tabellen in kleinere, zusammenhängende Tabellen → bessere Organisation von Daten und effizienteres Abfragen möglich
                - Einfache Zusammenführung von Inhalten über Join-Abfragen
                - Problem: Änderung von Daten über mehrere Relationen erfordert komplexe Anwendungslogik und Nutzung von Transaktionen
                - Web-Anwendungen: Nutzung objekt-orientierter oder anderer verschachtelter Datenstrukturen (z.B. JSON) und Arbeit mit komplexen Objekten (z.B. Nutzerprofil)
- 1990er: objekt-orientierte Datenmodelle oder XML
    - Erweiterung der Funktionalität relationaler Systeme
        - Objekt-relationale Funktionen (z.B. Vererbung)
        - XML-Datentyp mit entsprechender Abfragemöglichkeiten
- 2000er: kostengünstige Skalierbarkeit und Fehlertoleranz bei stark verteilten Anwendungen gewinnt mit steigender Popularität von großen Internetplattformen (Amazon, Ebay und Google) an Bedeutung

- Entwicklung von **NoSQL-Systeme**n
    - Ein für Anwendungen besser passendes Datenmodell als SQL
    - Direkte Unterstützung eines verteilten Daten-Managements über mehrere Server-Knoten → Fehlertoleranz gegenüber Netzwerkpartitionierungen
    - Nach CAP-Theorem: Wahl zwischen jederzeit konsistenten Datenbeständen oder Verfügbarkeit → Verfügbarkeit
    - Zwischenzeitlich inkonsistente Datenbestände als Folge (Eventual Consistency)
    - Datenmodelle der NoSQL-DBMS
        - Key-Value
            - Datenobjekte unter Schlüssel gespeichert
            - Zugriff auf Datenobjekte über Schlüssel oder auch einfache Abfragemechanismen möglich
            - Datenobjekte haben häufig baumartige Struktur (z.B. JSON-Dokument)
        - Dokumenten-orientiert
            - Ablage von Daten als JSON-Dokumente
            - System unterstützt weitergehende Abfragemöglichkeiten über Struktur der Dokumente
            - Bsp.: MongoDB als populärstes NoSQL-System
        - Wide Column
            - Dem relationalen Modell sehr ähnlich → vergleichbare Abfragemöglichkeiten wie SQL
            - Dynamische Anpassung der Spalten in Datensätzen möglich → nicht alle Datensätze müssen gleiche Struktur haben
        - Graph-orientiert
            - Abspeicherung von Graphen mit komplexen Knoten und Kanten möglich

        - Knoten und Kanten können verschiedene Typen und Attribute haben
        - Mathematische Eigenschaften von Teilgraphen (z.B. Konnektivität, kürzeste Wege) testen und nach bestimmten Mustern im Graphen suchen mithilfe von Abfragen
    - Vor- und Nachteile
        - Meist nicht erforderlich ein Schema für Daten zu definieren → direkte Nutzung möglich (Achtung: gewisse Modellierung oder Strukturierung von Daten unausweichlich)
        - Mehr Flexibilität und Skalierbarkeit
        - Wanderung eines Teils der Logik zur Überprüfung der Datenstruktur oder Integrität neuer Daten vom DBMS in Applikation → Erhöhung von Komplexität der Anwendungen und Implementierungsaufwand
- Heute
    - Meiste Systeme in einer kostenlosen Open-Source-Variante
    - Kostenpflichtige „Enterprise Editions“ für weitere Funktionalität

Data Engineering

Data Engineering ist in der Fachliteratur nicht genau definiert und wird oft im Kontext der Begriffe „Data Management“ und „Information Engineering“ verwendet.

- Datenmanagement ist ein ganzheitliches Konzept zum Umgang mit digitalen Daten, das alle Schritte vom Erheben, über das Speichern und die Verarbeitung bis hin zur Archivierung und Löschung umfasst.
- Information Engineering ist ein Ansatz, der darauf abzielt, Informationssysteme effektiv zu entwickeln, zu implementieren und zu verwalten, um die Geschäftsprozesse zu unterstützen.

Beispiele für Vorgehensmodelle in der Datenanalyse, die verdeutlichen, womit sich das Data Engineering beschäftigt.

- KDD: Knowledge Discovery in Databases
    - Verständnis des Problems: Identifikationder Fragestellung, die durch die Analyse gelöst werden soll.

    - Datenauswahl : Auswahl der relevanten Datenquellen, die für die Analyse verwendet werden sollen
    - Datentransformation: Umwandlung der vorverarbeiteten Daten in ein für die Analyse geeignetes Format.
    - Datenmining: Anwendung von Datenmining-Techniken, um Muster, Trends und Wissen aus den vorverarbeiteten Daten zu extrahieren.
    - Musterbewertung: Bewertung der extrahierten Muster und Trends hinsichtlich ihrer Relevanz für die gestellte Fragestellung.
    - Wissensdarstellung: Darstellung des extrahierten Wissens in einer für die Entscheidungsfindung verständlichen Form.
    - Wissensnutzung: Integration des extrahierten Wissens in den Entscheidungsprozess der Organisation.
- **CRISP-DP**: Cross Industry Standard Process for Data Mining
    - Geschäftsverständnis: Die Klärung des Umfangs und die Festlegung eines vorläufigen Plans zur Erreichung der Geschäftsziele,
    - Datenverständnis: Die Identifizierung von Datentypen, die Bewertung der Datenqualität und das Verständnis von Beziehungen innerhalb der Daten.
    - Datenpräparation: Das Reinigen, Transformieren und Auswählen von Daten, um einen geeigneten Datensatz für das Modellieren zu erstellen
    - Modellierung: Die Auswahl geeigneter Modellierungstechniken, die Identifizierung des am besten geeigneten Modells zur Erreichung der Geschäftsziele.
    - Bewertung: Das Testen der Modelle an unabhängigen Datensätzen und die Sicherstellung, dass die Ergebnisse gültig und zuverlässig sind.
    - Bereitstellung: Die Integration des Modells in Geschäftsprozesse
- Data Engineering beschäftigt sich vor allem mit den **Aufgaben , die vor der eigentlichen Datenanalyse stattfinden**.
    - Domain Understanding: Ein detailliertes Verständnis der Daten entwickelt sich nur durch ein Verständnis der Prozesse im Unternehmen, die Daten produzieren und konsumieren.
    - Die Formalisierung des Verständnisses über die Daten in einem Datenmodell: Die Formalisierung des Verständnisses über die Daten und die Erkennung von Verknüpfungen und Regeln ermöglicht die Beschreibung einer Struktur für neu zu erfassende Daten.
    - Die Aufbereitung und Integration von Daten: Die Umwandlung der Datenmenge in das gewünschte Format, erfolgt mithilfe verschiedener Methoden und Werkzeuge.
    - Die Definition einer effizienteren Daten-Management-Architektur: Eine Daten-Management-Architektur ermöglicht die Zusammenführung und

Verknüpfung von Daten in einem Unternehmen bei einer intensiven Kommunikation mit verschiedenen Unternehmensbereichen

KW 49, Thema

http://jbusse.de/dsci-101/dsci-101-quellen.html#term-Barton-Kokoev-2021

Barton, T., Kokoev, A. (2021). Text Mining bei einer wissenschaftlichen Literaturauswertung: Extraktion von Schlüsselwörtern zur Beschreibung von Inhalten. In: Barton, T., Müller, C. (eds) Data Science anwenden. Angewandte Wirtschaftsinformatik. Springer Vieweg, Wiesbaden. https://bibaccess.fh-landshut.de:2188/10.1007/978-3-658-33813-8_11

- 11.3 Extraktion von Schlüsselwörtern
    - Rapid Automatic Keyword Extraction (RAKE)
- 11.4 Extraktion von Schlüsselwörtern für eine Literaturauswertung zu „Explainable AI“

Text-Mining bei einer wissenschaftlichen Literaturauswertung

2. Beispiel: Explainable Artificial Intelligence

1. Was ist Explainable AI?

- KI-Systeme, die ihre Entscheidungsfindung erklären können

2. Verstehen

- „Intellektuelle Erfassung des Zusammenhangs“
- Informationen filtern & gruppieren
- Bsp.: Voraussetzungen, Aktionen, Ziele, ...

3. Erklären

- „Die Ursachen eines beobachteten Sachverhaltes durch eine sprachliche Darlegung seiner logischen und kausalen Zusammenhänge verständlich zu machen“
- Lösung sprachlich korrekt umsetzen und wiedergeben

4. Anwendung in der Gesellschaft

- Positive Reaktion im technischen & wissenschaftlichem Bereich?
- Negative Reaktion in der Gesellschaft bzw. bei Privatpersonen?

Extraktion von Schlüsselwörtern: Eine Einführung in Rapid Automatic Keyword Extraction (RAKE)

- Was ist RAKE ? Rapid Automatic Keyword Extraction auch bekannt als RAKE ermöglicht das zusammenfassen eines Textes mittels Schlüsselwörtern.
    - •Textanalyse und Informationsextraktion
    - •Suchmaschinenoptimierung (SEO)
    - •Dokumentensummarisierung
    - •....Vielen anderen bereichen
- Schlüsselwortextraktion mit RAKE
    - Schlüsselwörter beschreiben prägnant den Inhalt, unabhängig von Sprache und Domäne
    - 1. Anwärter für Schlüsselwörter bestimmen
        - • Der Text wird zuerst in einzelne Wörter oder Phrasen aufgeteilt, ein Vorgang, der als Tokenisierung bekannt ist.
        - • Häufig verwendete Wörter wie "und", "die", "ist" usw., die als Stoppwörter bezeichnet werden, werden entfernt. Diese Wörter werden in der Regel bei der Schlüsselwortextraktion ignoriert, da sie nicht wesentlich zum Gesamtverständnis beitragen.
        - • Die verbleibenden Wörter oder Phrasen werden als potenzielle Schlüsselwortkandidaten betrachtet.
        - • Verwendung von Stoppwörtern und Trennzeichen zur Aufteilung des Dokuments in Wörter.
        - • Eine Sequenz von benachbarten Wörtern ohne Irrelevanz wird als Schlüsselwortanwärter betrachtet.
        - • Schlüsselwortanwärter : Schlüsselwortanwärter sind Wörter oder Phrasen, die im Rahmen eines Algorithmus zur Schlüsselwortextraktion als potenzielle Schlüsselwörter betrachtet werden.
    - 2. Kennzahl für Schlüsselwörter ableiten
        - Jeder Kandidat wird anhand seiner Häufigkeit im Text und seines Vorkommens in Verbindung mit anderen Wörtern bewertet. Die Idee ist, dass wichtige Schlüsselwörter wahrscheinlich häufig auftreten und in sinnvoller Nähe zu anderen Wörtern stehen.
        - RAKE leitet die Kennzahl K(w) für Schlüsselwortanwärter ab.
        - Eine Matrix wird erstellt, wobei Zeilen und Spalten durch Schlüsselwortanwärter gebildet werden.

        - Durch Matrixelemente werden Worthäufigkeit freq(w) und Wortmaß deg(w) ermittelt. K(w) = deg(w) / freq(w)
    - 3. Schlüsselwörter festlegen
        - • Endgültige Schlüsselwörter werden ausgewählt
        - • -- > die Schlüsselwortanwärter mit den höchsten Werten für die Kennzahl K(w)
        - • Diese repräsentieren die bedeutendsten Wörter zur optimalen Beschreibung des Dokumentinhalts.
- Fazit
    - RAKE ist eine effektive Methode zur automatischen Extraktion von Schlüsselwörtern.
    - Durch klare Strukturierung von Schlüsselwortanwärtern und Ableitung einer aussagekräftigen Kennzahl.
    - Sprach- und domänenunabhängigkeit ermöglicht RAKE eine effiziente Analyse und Zusammenfassung von Dokumentinhalten.

KW 50, Thema

http://jbusse.de/dsci-101/dsci-101-quellen.html#term-Peuker-Berton-2021

Peuker, A., Barton, T. (2021). Empfehlungssysteme und der Einsatz maschineller Lernverfahren. In: Barton, T., Müller, C. (eds) Data Science anwenden. Angewandte Wirtschaftsinformatik. Springer Vieweg, Wiesbaden. https://bibaccess.fh-landshut.de:2188/10.1007/978-3-658-33813-8_6

Klassifikation von ES

6.1 Einleitung

- Nutzer-Objekt-Matrix

6.2 Kollaborative Empfehlungssysteme

- 6.2.1 Ansätze
    - nutzerbasierten Ansatz
    - objektbasierten Ansatz
- 6.2.2 Methoden
    - Cosinusähnlichkeit
    - Clustering
    - Klassifizierung

6.3 Inhaltsbasierte Empfehlungssysteme

- 6.3.1 Ansatz
    - Objektprofile
    - Nutzerprofile
    - Filterkomponente
- 6.3.2 Methoden
    - TF-IDF (Term Frequency times Inverse Document Frequency)

6.4 Weitere Konzepte

- 6.4.1 Demografische Empfehlungssysteme
- 6.4.2 Wissensbasierte Empfehlungssysteme
- 6.4.3 Hybride Empfehlungssysteme

6.5 Aktuelle Entwicklungen

KW 50 Demographische Empfehlungssysteme

Funktionsweise

- Einordnung in Gruppen
- Basiered auf demographischen Daten
- Empfehlungen basierend auf Gruppenzuordnung

Vorteile

- Einfache Erhebung der Daten
- Erforschung von Nischen
- Effizient bei großer Nutzerzahl

Nachteile

- Erfordern persönlicher Daten
- Keine Empfehlungen außerhalb eingeordneter Gruppe

Empfehlungssyteme und der Einsatz maschineller Lernverfahren

Aktuelle Entwicklungen

Konzepte und Ansätze als Anwendungsgrundlage

- Stetige Weiterentwicklung der anzuwendenden Methoden
- Besondere Aufmerksamkeit für Methoden im Bereich des maschinellen Lernens

Methoden im Bereich des maschinellen Lernens

- Untersuchung des Einsatzes von Methoden für Empfehlungssysteme

- Bayes'sche Methoden und Entscheidungsbäume für Empfehlungsgenerierung
- Betonung auf geringere Komplexität dieser Methoden

Entwicklung in den letzten Jahren

- Zunehmender Einsatz von Deep Learning-Methoden
- Erfolgreiche Beispiele von Unternehmen wie Google, Facebook und Amazon
- Veröffentlichung von Amazon's Deep Learning Framework DSSTNE unter Open-Source-Lizenz

Forschungsdiskussion zu Deep Learning

- Vergleich von Deep Learning mit herkömmlichen Methoden in Wettbewerben
- Häufige Übertreffen von bestehenden Verfahren in Bezug auf Performance oder Vorhersagegenauigkeit
- Beobachtung, dass herkömmliche Methoden in den meisten Fällen erfolgreich sind
- Mögliche Ursachen, wie Laufzeitverhalten und Datenvolumen für das Training von neuronalen Netzen

Training von neuronalen Netzen und Datenvolumen

- Unterschiede zwischen Forschung und Wettbewerben hinsichtlich Datenvolumen
- Ressourcenschonendere Methoden in Wettbewerben aufgrund großer Datensätze
- Lange Rechenzeit für das Training von neuronalen Netzen bei großen Datenvolumen
- Unternehmen verfügen typischerweise über ausreichende Rechenleistung für effizientes Training

Bedeutung der herkömmlichen Methoden

- Betonung der weiterhin aktuellen Bedeutung herkömmlicher Methoden des maschinellen Lernens
- Wichtigkeit auch im Hinblick auf große Datensätze und Ressourcenanforderungen für neuronale Netze

Ansatz der Inhaltsbasierten Empfehlungssysteme

Profilerstellung

- Objektprofile
  - Enthalten charakteristische Eigenschaften
- Userprofile

- Enthalten Nutzereingaben

Profilabgleich

Inhaltsbasierte Empfehlungssysteme:Methoden

Verwendung des Vektorraummodells

- Büchern Eigenschaften
- Textdokumente:Tokenisierung,Stemming,Entfernung von Stoppwörtern
- Präsentation von Textdokumenten
    - Vektorraummodell:TF-IDF-Gewichtungsfaktor
    - Worteinbettung:Wörter einbetten

Erstellen eines Benutzerprofils

- Profilerstellung:Feedback zu Objekten
- Vorhersage von Interessen:Kosinusähnlichkeit von Vektoren

Anwendung in der Praxis

- Verständnis der Prinzipien von Empfehlungssystemen
- Wirksamkeit in realen Szenarien:Online-Shops,Streaming-Dienste,Informationsplattformen

Methoden von Empfehlungssystemen für Informationsinhalte

- Ein wichtiges Personalisierungstool
- Verbesserung der Benutzererfahrung

Hybride Empfehlungssysteme

Hybride Empfehlungssysteme

- Kombination aus inhaltsbasierten und kollaborativen Filtermethoden
- Überwindung der Grenzen einzelner Algorithmen
- Nutzung der Vorteile verschiedener Ansätzen

Vorteile der HEs

- Verbesserte Genauigkeit
- Anpassungsfähigkeit
- Robustheit

Nachteile der HEs

- Ressourcenintensiv
- Komplexität

- Mangelnde Erklärbarkeit

Beispiele für HEs

- Amazon
- Spotify
- Netflix

Netflix Empfehlungssysteme

- Ähnlichkeiten mit anderen Mitgliedern
- Nutzerinteraktionen (z.B angesehene Titel, Bewerbung)
- Details zu Titeln (Genre, Schauspieler, Erscheinungsjahr)
- Nutzungsdauer
- Verwendete Geäte

Kollaborative Empfehlungssysteme

Ansätze

- Nutzbasierter Ansatz: Ähnlichkeiten zwischen Nutzern anhand der Korrelation ihrer Bewertungen berechnet werden
- Objektbasierter Ansatz: Ähnlichkeiten zwischen Objekten anhand der Korrelation des Nutzerfeedbacks berechnet werden

Methoden

Speicherbasierte Methoden

- Die Aehnlichkeit zwischen Nutzern oder Objekten berechnen
- Gesamte Nutzer-Objekte Matrix unter Verwendung der Cosinusähnlichkeit nutzen:

Modellbasierte Methoden

-
  - Ein statistisches Modell generieren, mittels Methoden des maschinellen Lernens.
- Typische Methoden
  - Clustering: Gruppierung der Daten in verschiedene Cluster nach „Ähnlichkeit“
  - Klassifizierung: Ein Modell mittels eines Datensatzes trainiert wird. Hier wird der Bayes’scher Klassifikator verwendet

KW 51, Thema

http://jbusse.de/dsci-101/dsci-101-quellen.html#term-Hammesfahr-Spott-2021

Hammesfahr, J., Spott, M. (2021). Identifikation relevanter Zusammenhänge in Daten mit maschinellem Lernen. In: Barton, T., Müller, C. (eds) Data Science anwenden. Angewandte Wirtschaftsinformatik. Springer Vieweg, Wiesbaden. https://bibaccess.fh-landshut.de:2188/10.1007/978-3-658-33813-8_12

Identifikation relevanter Zusammenhänge in Daten mit maschinellem Lernen (kW51)

Einleitung:

- Bsp.: Bereich der Telekommunikation
- Ziel: Reduzierung des Aufwandes ohne Kunden zu verärgern
- Nutzung von Daten
    - Ansätze für Fehlerreduktion
- Identifikation von Zusammenhängen durch:
    - Subgroup Discovery
    - Lernverfahren für Assoziationsregeln
    - Problem: zu große Anzahl an Zusammenhängen
- ZIEL: Reduzierung von Zusammenhängen auf handhabbare Größe, ohne interessante Zusammenhänge zu verlieren

Fachliche Problemstellung:

- Datenbasis
    - fachliche Domäne eines Fahrzeugherstellers
    - anonymesierte, reale, betriebliche Daten
    - keine Veränderung der statistischen Struktur der Zusammenhänge

Ansätze zur Reduzierung von Regelmengen:

- Rule Learning:
    - beschäftigt sich mit dem Thema 'von gegebener Datenbasis interessante Regeln ableiten'
- Descriptive Rule Discovery:
    - wie individuelle interessante Muster in Daten extrahiert werden können
- Association Rule Discovery:
    - Generierung von Assoziationsregeln in einer Datenbasis
- Subgroup Discovery:

    - Ableitung interessanter Zusammenhänge
    - mit Bezug auf eine festgelegte Eigenschaft der Individuen einer Datenbasis

Gütebestimmung von reduzierten Regelmengen:

- Ganzheitlichkeit
- Komplexität
- Interessantheit
- Redundanzfreiheit

Kombinationssystematik:

- mögliche Beziehungen zw. zwei Regeln
    - eine Teilmengenbeziehung
    - keine Teilmengenbeziehung, eine Schnittmenge
    - keine Schnittmenge

Ableitung von fünf Schritten:

- Entfernung reiner Redundanz
- Approximation ähnlicher Regeln mit einer Teilmengenbeziehung
- Approximation ähnlicher Regeln mit einer dominanten Schnittmenge
- Anwendung von Beschränkungen
- Selektion einer interessanten Regelmenge

Ergebnisse:

- Implementierung in eine Programmiersprache
- Reduzierung der Regelmengen anhand von der Systematik
- wenige Regeln mit wenig Redundanz selektieren
- zwei reduzierte Regelmengen, die sich bzgl. der Evaluationsgrößen als gut bewerten lassen
- manuelle Auswertung durch Experten leicht handhabbar
- einzelne Regeln sind unterschiedlich
    - gezielte Betrachtung relevanter Faktoren

Zusammenfassung:

- Ziel
    - interessante Zusammenhänge zw. Produktkonfigurationen und Produktfehlern
    - Reduzierung Anzahl der Zusammenhänge auf ein handhabbares Maß ohne Informationen zu verlieren
- Ergebnis

    - Reduzierung der 165.720 Zusammenhänge auf 2 mögliche Regelmengen mit 81 und 24 Regeln
    - kleine Anzhal an Regeln ermöglicht die Zusammenhänge sequenziell durchzugehen
    - fachliche Bewertung durch Experten
    - ob Erkenntnisse für eine Verbesserung der Produktqualität gewonnen werden kann
- Weg
    - Einbringung von Kontextwissen der Experten für optimale Komprimierung

Gütebestimmung von reduzierten Regelmengen

Allgemeines

- Messgrößen werden zur Bewertung einer Reduktion einer Regelmenge in Bezug auf die ausgehende Zielsetzung benötigt

qualitative Eigenschaften zur Bewertung von Mustern aus der Literatur

- zur Gruppierung von Kennzahlen im Rahmen der Subgroup Discovery
    - Komplexität
    - Generalität
    - Genauigkeit
    - Interessantheit
- subjektive Messgrößen zur Bewertung von Interessantheit
    - Redundanz
    - Neuheit
    - Unerwartbarkeit
    - Nützlichkeit
    - Aktionsfähigkeit

aus der Literatur ermittelte qualitative Eigenschaften

Ganzheitlichkeit

- reduzierte Version der Datenbasis soll keine relevanten Informationen verlieren

Komplexität

- wird durch die Anzahl der Regeln bestimmt

Interessantheit

- Messung erfolgt anhand durchschnittlicher Werte der jeweiligen Qualitätskennzahlen einer Regel

- Ableitung der Tendenz einer einfachen Kennzahl erfolgt über eine Rang-Funktion, anschließend wird Durchschnitt des besten und schlechtesten Ranges gebildet
- soll gegenseitige Bekräftigung ähnlicher Qualitätsfunktionen verhindern

Redundanzfreiheit

- viele redundante Regeln enthalten als Ganzes relativ zur Regelmenge wenig neue Informationen
- Kennzahl für Redundanz ist die durchschnittliche Abdeckung eines Datensatzes durch eine Regel (entspricht Expected Cover Count)

Fachliche Problemstellung

Datenbasis: Fahrzeughersteller

Urprung der Daten

- Fiktive Fahrzeughersteller
- Anonymisierung der Begrifflichkeiten

Wichtiger Erfolgsfaktor

- Kundenzufriedenheit

Einflussfaktoren auf Kundenzufriedenheit

- Qualität der produzierten Fahrzeuge
    - Messung durch FAULT_RATE
    - Verschieden FAULT_TYPE

Merkmale zu identifizierung der Ursachen

- AGE
- DEALERSHIP
- CUSTOMER_TYPE
- COUNTRY
- GEO_TYPE
- MODEL
- USER_CUSTOMIZED

Aggregation der Daten

- CAR_COUNT
- FAULT_COUNT
- Fehlerrate

Alternative Herangehensweise zur Identifizierung interessanter Zusammenhänge

Visuelle Exploration der Datenbasis

- Exemplarische Abbildung für Fahrzeugmodelle(Bsp. Abb. 12.2)
    - Hopper
    - Quantum
    - Ultima
- Boxplots der Fehlerraten
- Unterschiede zwischen Modelle und Fehlertypen
- Notwendigkeit einer übersichtlichen Zusammenfassung

Werteausprägungen der Merkmale

AGE

- In Warranty
- Out of Warranty

DEALERSHIP

- Franchise
- Re-import
- Branch
- Used Car Dealer

CUSTOMER_TYPE

- Other
- Private
- Corporate

COUNTRY

- Portugal
- Germany
- ....(Usw.)

GEO_TYPE

- Suburban
- Urban
- Village
- ....

MODEL

- Opal(Abb. 12.1)
- Hopper (Abb. 12.2)

FAULT_TYPE(Abb. 12.2)

- Air conditioning
- Break Fluid
- ....

KW_51 / Kapitel 5. Kombinationssystematik

1.Entfernung reiner Redundanz

- Ziel: Identifikation und Eliminierung von Teilmengenbeziehungen
- Schritte:
  - Identifizierung von Teilmengenbeziehungen
  - Anwendung von Closed Non-Derivable Itemsets zur Entfernung redundanter Regeln

2. Approximation ähnlicher Regeln mit einer Teilmengenbeziehung

- Ziel: Entfernung von Regeln mit überlappendem Informationsgehalt
- Schritte:
  - Filterung durch Positive Improvement
  - Zusammenfassung ähnlicher Regeln durch Negative Replacement
  - Verwendung von Condensed Itemsets für die finale Zusammenfassung

3.Approximation ähnlicher Regeln mit einer dominanten Schnittmenge

- Ziel: Reduzierung von Redundanz durch Überlappungskomprimierung
- Schritte:
  - Anwendung von Subgroup Suppression
  - Berücksichtigung verschiedener Qualitätskennzahlen bei der Ergebniszusammenführung

4. Anwendung von Beschränkungen

- Ziel: Filtern der verbleibenden Regelmenge
- Schritte:
  - Festlegung von Mindestsupport und Mindestkonfidenz
  - Anpassung der Beschränkung "Minimal Improvement" unter Berücksichtigung bereits gefilterter Regeln

5. Selektion einer interessanten Regelmenge

- Ziel: Auswahl diverser Regeln für umfassenden Einblick
- Schritte:
  - Anwendung von verschiedenen Selektionsstrategien

- Auswahl von Regeln, die vielfältige Einblicke bieten

Empfehlungssysteme

Inhaltsbasierende Empfehlungssysteme

- 3 wesentliche Schritte:
  - Präferenzen ermitteln
  - für jedes Objekt ein Profil anlegen
  - abgleichen
- Methoden
  - Vektor-Raum-Modell
  - Word Embedding

Kollaborative Empfehlungssysteme

- 2 Ansätze
  - Nutzerbasierter Ansatz
  - Objektbasierter Ansatz
- Methoden
  - Speicher-basiert
  - Modellbasiertes-kollaboratieves-Filtern
  - Clustering
  - Bayes'scher Klassifikation

demografische Empfehlungssysteme

- nutzt vordefinierte Stereotypen
- nutzt auch Nutzer-Objekt-Matrix

hybride Empfehlungssysteme

- gleicht Nachteile des einen, mit Vorteilen das anderen Empfehlungssystem aus