Identifikation typischen Benutzerverhaltens in digitalen Studienformaten

Bachelorarbeit zur Erlangung des akademischen Grades Bachelor of Science
Berliner Hochschule für Technik · Fachbereich VI · Informatik und Medien

AUTOR
Werner Breitenstein
Matrikelnr.: 866059

BETREUER
Prof. Dr. Petra Sauer

GUTACHTER
Prof. Dr. Heike Ripphausen-Lipa

ABGABE
dd.mm.2022

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Quellcodeverzeichnis

Zusammenfassung

...

Abstract

...

1 Einleitung

Ziel- und Endpunkt der Arbeit ist die detaillierte Analyse und Dokumentation des IST-Zustands. Es werden weder Prognosen abgeleitet noch Empfehlungen gegeben.

...

2 Grundlagen

In diesem Kapitel werden die theoretischen Grundlagen dieser Arbeit beleuchtet und mithin wichtige Informationen zur angewandten Methodik, zu technischen Mitteln und zu dem zu untersuchenden Gegenstand bereitgestellt.

Ausgehend von in der Wissenschaft und in der Industrie seit langer Zeit anerkannten standardisierten Vorgehensmodellen wie dem *KDD – Knowledge Discovery in Databases Process* – (Fayyad, Piatetsky-Shapiro & Smyth, 1996) bzw. dem etwas jüngeren *CRISP-DM – Cross Industry Standard Process for Data Mining* – (Shearer, 2000) wird zunächst das im Rahmen dieser Arbeit praktizierte Analyseverfahren skizziert sowie die wesentlichen Grundlagen der explorativen Datenanalyse und der Visualisierung von Daten beschrieben.

Im folgenden zweiten Abschnitt werden die im Zuge der zahlreichen praktischen Untersuchungen eingesetzten Werkzeuge und Technologien vorgestellt.

Unter verschiedenen Aspekten wird abschließend die Datenbasis betrachtet und präsentiert. So werden hier die Daten u. a. durch Angaben zu ihrer Herkunft, ihrer Zusammensetzung und ihrer Qualität zum einen formal beschrieben. Statistische Abfragen sowie erste Visualisierungen z.B. zu bestehenden Mengengerüsten geben hier aber auch bereits interessante Einblicke in Struktur und Inhalt der Daten.

2.1 Theorie

Der Wunsch, Wissen aus Daten zu extrahieren, ist nicht nur sinnstiftend für diese Arbeit. Vielmehr ist er in der heutigen Informationsgesellschaft, in der viele erfolgreiche Geschäftsmodelle wie die der Big Five[1] gerade auf einer intelligenten wirtschaftlichen Verwertung dieser Ressource beruhen, nahezu allgegenwärtig.

[1] Die Bezeichnungen *The Big Five* oder auch *GAFAM* gelten den fünf größten globalen Technologieunternehmen: Google, Apple, Facebook, Amazon und Microsoft: Statista, 01/2020

Aber nicht nur Google, Apple und andere haben früh erkannt, dass Daten gerade auch mit Blick auf ihr expansives Wachstum eine sehr ergiebige Quelle wertvoller Informationen[2] darstellen, sondern auch die Wissenschaften.

Diese letzteren waren es, die schon in den 1980er Jahren damit begonnen haben, Daten nicht nur sporadisch auf interessante Muster hin zu untersuchen, sondern unter dem Begriff *Data Mining* und später auch *Data Analytics* strategisch sinnvolle und allgemeingültige Prozesse zu etablieren (Runkler, 2020).

2.1.1 Standardisierte Vorgehensmodelle der Datenanalyse

Neben organisatorischen und wirtschaftlichen Erwägungen waren und sind es auch einfach faktische Gegebenheiten, die die Notwendigkeit der Standardisierung und Automatisierung von Analyseprozessen früh verdeutlichte und über die Jahre viele Experten zu entsprechenden Lösungsansätzen motivierte.

Denn wie Runkler (2020) und andere schreiben, ist die Datenanalyse ein stark interdisziplinärer Prozess, bei dem je nach Kontext oft mehrere Personen aus ganz unterschiedlichen Fachbereichen zusammenkommen. Damit liegt es auf der Hand, dass hier in einem äußerst heterogenen Umfeld von Experten, u. a. für Statistik, für maschinelles Lernen oder für Datenbanksysteme, die Orientierung an einem klar strukturierten Verfahren die Zusammenarbeit erheblich vereinfacht.

Konkrete wirtschaftliche Vorteile durch Zeit- und Kosteneinsparungen und die größere Objektivität bei der Durchführung der Analyse werden von Fayyad et al. (1996) als wichtige weitere Motive genannt. Schon im Jahr 1996 erkannten sie aber auch das Problem des *Data Overload* in manchen Bereichen der Forschung und sie wiesen darauf hin, dass ein organisierter Prozess unbedingt erforderlich ist, um die faktische Durchführbarkeit einer Datenanalyse überhaupt zu gewährleisten.

[2] Siehe hierzu die geschätzten Mengen der E-Mails, WhatsApp-Nachrichten oder YouTube-Uploads, die jede Minute allein im Internet entstehen bzw. verarbeitet werden: Statista, 06/2021

KDD – Knowledge Discovery in Databases Process

Der *Knowledge Discovery in Databases Process* (*KDD*), wie er von Fayyad et al. (1996) geprägt wurde, beschreibt einen umfassenden Datenanalyseprozess, in dessen Kern Verfahren des Data Mining zur Anwendung kommen.[3]

Die folgende Übersicht veranschaulicht die fünf verschiedenen Phasen des KDD – *Selektion*, *Vorverarbeitung*, *Transformation*, *Data Mining*, *Interpretation/Evaluierung* –, die, wie durch die gestrichelten Pfeile angedeutet, bei einer Analyse in vielen Fällen auch wiederholt durchlaufen werden müssen, bis tatsächlich ein aussagekräftiges Ergebnis vorliegt.

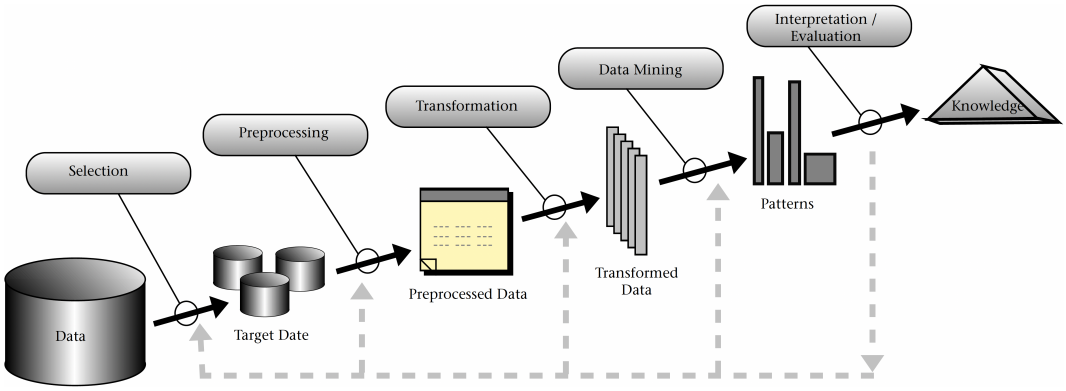

Abbildung 1: Phasen des KDD-Prozesses. Original von Fayyad et al. (1996).

Über die genaue Zuordnung und Differenzierung von Arbeitsschritten innerhalb der oben dargestellten Hauptphasen des KDD, gibt es in der Literatur verschiedene Meinungen. Azevedo & Santos (2008) ordnen diese wie folgt ein:

1. *Selektion:* Auswahl des relevanten Teils des Datenbestands, der als Gegenstand der Untersuchung geeignet erscheint.
2. *Vorverarbeitung:* Zusammenführung und Bereinigung der selektierten Daten, bei der u. a. falsche und inkonsistente Daten entfernt werden sollten.
3. *Transformation:* Überführung der Daten u. a. mittels Konvertierung von Datentypen, wodurch z. B. verschiedene Datumsformate vereinheitlicht werden.

[3] Unter dem folgenden Link findet sich dauerhaft die zitierfähige Version der im Text erwähnten Definition: Gabler Wirtschaftslexikon, Springer Gabler, 04/2022

4. *Data Mining:* Anwendung von Methoden und Algorithmen mit deren Unterstützung möglichst automatisch empirische Zusammenhänge aus der bereitgestellten Datenbasis extrahiert werden sollen.[4]

5. *Interpretation/Evaluierung:* Auslegung und Prüfung der gewonnenen Erkenntnisse, ggf. unterstützt durch Visualisierung extrahierter Muster.

CRISP-DM – Cross Industry Standard Process for Data Mining

Der *Cross Industry Standard Process for Data Mining* (*CRISP-DM*) ist ein auf Basis eines ehemals durch die EU geförderten Projekts entstandenes anwendungs- und branchenunabhängiges Vorgehensmodell für das Data Mining.

Konzipiert und entwickelt wurde das Vorhaben in den Jahren 1996 bis 2000 durch ein Konsortium namhafter Industrieunternehmen, der CRISP-DM Special Interest Group, der damals u. a. Daimler-Benz, NCR und ISL angehörten. Ihr Ziel war es, für Data Mining-Projekte ein nicht-proprietäres Standard-Prozessmodell zu etablieren, das konkret als Blaupause dienen kann, um Datenbestände z. B. nach interessanten Mustern und Trends zu durchsuchen (Shearer, 2000).

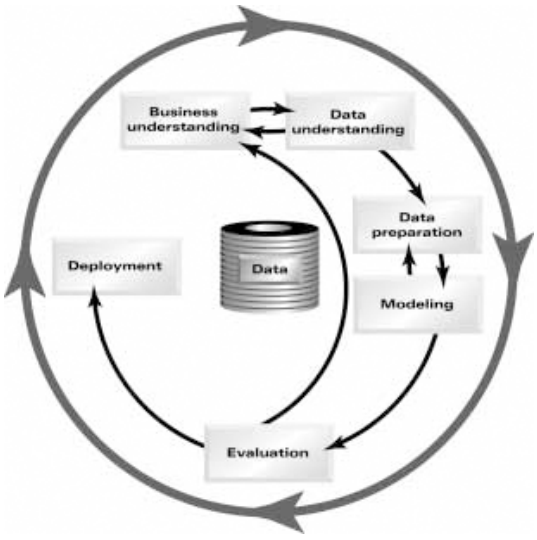

Abbildung 2: Phasen des CRISP-DM. Original von Shearer (2000).

[4] Unter dem folgenden Link findet sich dauerhaft die zitierfähige Version der im Text erwähnten Definition: Gabler Wirtschaftslexikon, Springer Gabler, 04/2022

Wie in der obigen Abbildung ersichtlich, umfasst der CRISP-DM insgesamt sechs Phasen, die hiernach in einem normalen Data Mining-Projekt zu durchlaufen sind. Ähnlich wie beim KDD können sich verschiedene Phasen dabei wiederholen oder es wird auch ein Springen zwischen den einzelnen Phasen erforderlich.

Die Ziele und Aufgaben der einzelnen Phasen des CRISP-DM lassen sich nach Shearer (2000) folgendermaßen kurz zusammenfassen:

1. *Geschäftsverständnis:* Beschreibung übergeordneter Ziele, Anforderungen und Beschränkungen; Definition von Strategien, Aufgaben und Methoden.
2. *Datenverständnis:* Sammlung und Beschreibung der Rohdaten; Prüfung und Bewertung der Datenqualität; Feststellung von Datenmängeln.
3. *Datenaufbereitung:* Auswahl, Zusammenführung, Bereinigung und Transformation der Daten zur Erstellung des zu untersuchenden Datenbestands.
4. *Modellierung:* Auswahl und Anwendung geeigneter Modellierungstechniken; Erstellung von Tests; Bewertung und Optimierung von Modellen.
5. *Evaluierung:* Bewertung der Analyseergebnisse und der genutzten Modelle; Prüfung des Gesamtprozesses; Ableitung nachfolgender Verfahrensschritte.
6. *Einsatz:* Aufbereitung und Vorstellung der gewonnenen Erkenntnisse; Ausarbeitung von Strategien und Maßnahmen zur Einführung und dauerhaften Verwendung;

Vergleich der standardisierten Vorgehensmodelle

Zum Abschluss dieses Kapitels über die standardisierten Vorgehensmodelle in der Datenanalyse soll hier noch einmal auf die Arbeit von Azevedo & Santos (2008) hingewiesen werden, die zum Ziel hatte die Gemeinsamkeiten und Unterschiede von KDD, CRISP-DM und SEMMA[5] miteinander zu vergleichen.

[5] Unter dem folgenden Link findet sich eine kurze Einführung zu SEMMA, das den übergeordneten Prozess für den SAS® Enterprise Miner™ darstellt: Introduction to SEMMA, SAS, 04/2022

Im Ergebnis bestätigt diese Vergleichsstudie die vollkommene Übereinstimmung von KDD und SEMMA, bzw. definiert SEMMA als praktische Implementation des älteren KDD-Prozesses, weshalb auch in dieser Arbeit auf eine Darstellung dieses Standardprozesses verzichtet wurde.

Im Vergleich von KDD und CRISP-DM gibt es dagegen erkennbare Unterschiede, die sich darin zeigen, dass der CRISP-DM die im KDD implizit enthaltenen vor- und nachgelagerten Stufen explizit als separate Teil des Prozesses ausführlich beschreibt. Weitere Abweichungen lassen sich feststellen bei der Zuordnung von Teilschritten innerhalb des *Data Understanding* und *Data Preparation*. Interessanterweise wird dies in dieser Studie nicht konsistent behandelt, und stimmt daher auch nur bedingt mit dem ursprünglich von Shearer (2000) skizzierten Prozess überein.

KDD	SEMMA	CRISP-DM
Pre KDD	----------	Business understanding
Selection	Sample	Data Understanding
Pre processing	Explore	
Transformation	Modify	Data preparation
Data mining	Model	Modeling
Interpretation/Evaluation	Assessment	Evaluation
Post KDD	----------	Deployment

Abbildung 3: KDD, SEMMA, CRISP-DM. Original von Azevedo & Santos (2008).

2.1.2 Angepasstes Vorgehensmodell für diese Arbeit

Die im vorausgegangenen Abschnitt präsentierten Vorgehensmodelle haben allesamt dasselbe Ziel: Sie wollen den äußerst vielfältigen Prozess einer Datenanalyse möglichst vollständig und genau in einem Standardverfahren abbilden und für den Anwender sinnvolle Handlungsempfehlungen formulieren.

Diese Verfahren sind also keineswegs verpflichtend. Sie sollen zur Orientierung dienen, aber es obliegt demnach stets dem Anwender je nach Anwendungskontext die standardisierten Verfahrensschritte auf die im konkreten Fall vorliegenden Anforderungen anzupassen (Shearer, 2000).

Grundzüge des verwendeten Vorgehensmodells

Im Hinblick auf die anstehenden Untersuchungen im Rahmen dieser Arbeit, wird das im weiteren Verlauf verwendete Vorgehensmodell – auf Basis des von Shearer (2000) beschriebenen CRISP-DM – wie folgt skizziert:

1. *Geschäftsverständnis:* Das Thema dieser Arbeit definiert gleichzeitig auch das übergeordnete Ziel, die *Identifikation typischen Benutzerverhaltens in digitalen Studienformaten*. Untergeordnete Ziele lassen sich mit Blick auf die Methodik und den Gegenstand der Untersuchung beschreiben. So gilt es, wie in der Einleitung zu dieser Arbeit beschrieben, mit Mitteln der explorativen Datenanalyse den Ist-Zustand studentischen Lern- und Kommunikationsverhaltens möglichst detailliert zu skizzieren und das jeweilige Vorgehen dabei verständlich und nachvollziehbar zu dokumentieren. Dazu bedarf es im Rahmen der eigentlichen Analyse neben der bestimmten Auswahl von Daten gerade auch der gezielten Entwicklung von Fragen, die geeignet sein könnten, das in den Daten verborgene Benutzerverhalten zu offenbaren und davon ausgehende neue Annahmen zu formulieren.

2. *Datenverständnis:* Ein fundiertes Verständnis über die Herkunft der zu untersuchenden Daten, deren Bedeutung und Qualität ist essentiell, um mögliche Zusammenhänge zu verstehen oder neues Wissen aus den Daten extrahieren zu können. Das nachfoldende Kapitel Datenbasis trägt diesem grundlegenden Erfordernis Rechnung und gibt detailliert Aufschluss über den Gegenstand der Untersuchung.

3. *Datenaufbereitung:* Im Fokus dieser Phase steht der konkrete Untersuchungsgegenstand. Dessen Bereitstellung vollzieht sich entsprechend der gegebenen Zielsetzung in mehreren Schritten. Zu nennen sind hier in erster Linie:

    - Datenauswahl: Die für die Untersuchung relevanten Daten sind nach Art und Umfang aus den Spalten und Zeilen der initial vorbereiteten Daten zu selektieren. Warum gewisse Daten relevant sind bzw. diese nicht in der Auswahl berücksichtigt werden, sollte begründet werden können.

- Datenbereinigung: Da die Daten initial keine falschen Werte aufweisen, entfällt naturgemäß eine entsprechende Korrektur. Gegebenfalls müssen aber fehlende Werte ergänzt werden, um bestimmte Abfragen sinnvoll durchführen zu können.
- Datentransformation: Für eine Untersuchung kann es erforderlich sein, zuvor aus den Daten ein neues Attribut abzuleiten, den Datentypen eines Attributs zu konvertieren oder auch weitere Datensätze zu ergänzen. Die Gründe hierfür sollten ebenfalls klar ersichtlich dokumentiert werden.

4. *Datenanalyse:*[6] Das Verfahren, das bei den eigentlichen Untersuchungen zur Anwendung kommen soll, orientiert sich an der Methodik der explorativen Statistik bzw. der explorativen Datenanalyse. Insbesondere durch geeignete visuelle Darstellungen[7] sollen in den Daten bemerkenswerte Strukturen und Zusammenhänge aufgezeigt werden, die zur Formulierung von Hypothesen anregen. Mögliche Darstellungsformen sind beispielsweise:

   - Balkendiagramm
   - Streudiagramm
   - Liniendiagramm

   Aufgrund komplexer Fragestellungen und Zwischenbewertungen sind bei der Analyse oft mehrere Anläufe nötig, um schließlich interessante Hypothesen generieren zu können. Gegebenenfalls muss auch die Frage selbst angepasst werden bzw. sind auch die Daten erneut aufzubereiten.

5. *Evaluierung:* Die Interpretation und die Bewertung von Analyseergebnissen vollzieht sich typischerweise im stetigen Wechsel mit der Optimierung der Methoden in der vorhergehenden Analysephase. Das Ziel ist dabei nur die Entwicklung einer Hypothese auf den erkannten Mustern oder Verbindungen in den Daten, nicht aber die Evaluierung der Hypothese selbst oder die Ableitung weiterer Verfahrensschritte aus einer gewonnenen Hypothese.

[6] Im weiteren Verlauf der Arbeit soll diese Phase vorzugsweise *Datenanalyse* genannt werden, da der Begriff Modellierung häufig die Anwendung komplexer Machine Learning Modelle impliziert.
[7] Siehe hierzu auch das nachfolgende Kapitel Formen der Datenvisualisierung

6. *Dokumentation:*[8] Erkenntnisse aus den Untersuchungen sind letztlich noch verständlich aufzubereiten und umfassend zu dokumentieren, so dass diese z. B. auch in einer neuen Studie zur Entwicklung von Kursempfehlungen genutzt werden könnten. Im Kapitel Ergebnisse werden dazu wichtige Erfahrungen aus dieser Arbeit zusammengefasst sowie bemerkenswerte Untersuchungsansätze und deren Resultate betrachtet bzw. miteinander verglichen.

Dieses Modell wird später bei der tatsächlichen Durchführung der Analyse (siehe das folgende Kapitel Analyse) erneut als Vorlage dienen und wie erwähnt in den Phasen *Datenaufbereitung*, *Datenanalyse* und *Evaluierung* je nach Anforderung auch mehrmals spezifisch angepasst werden müssen.

Die nachfolgende Grafik zeigt das in dieser Arbeit verwendete Vorgehensmodell mit den oben beschriebenen Phasen. Die nur im Rahmen der konkreten Analyse zu durchlaufenden Phasen sind dabei farblich hervorgehoben.

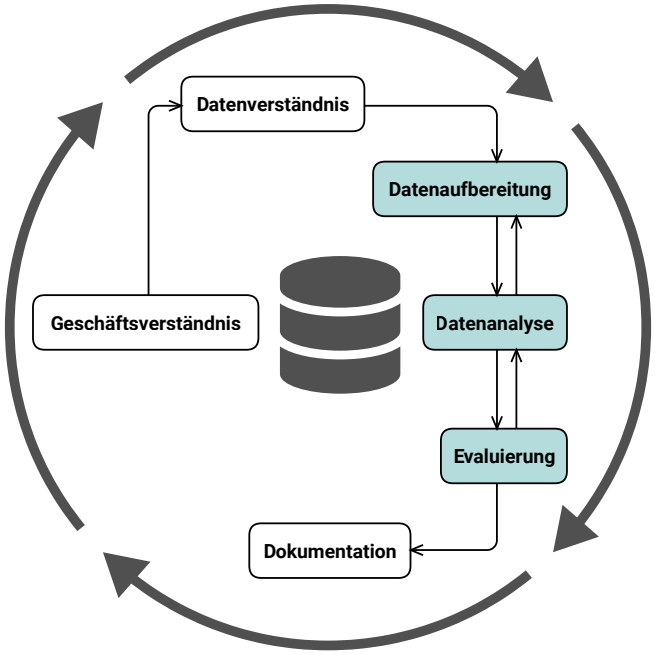

Abbildung 4: Phasen des verwendeten Vorgehensmodells.

[8] In dieser Arbeit soll diese Phase bevorzugt mit *Dokumentation* bezeichnet werden, da der Begriff Einsatz zu sehr auf die praktische Anwendung konkreter Untersuchungsergebnisse abzielt.

2.1.3 Explorative Datenanalyse

…

2.1.4 Formen der Datenvisualisierung

…

2.2 Technik

Hier finden sich Ausführungen zu den verwendeten Technologien, Tools, Libraries, etc.

...

2.3 Datenbasis

Gegenstand der Untersuchungen zu dieser Arbeit ist ein durch die *Virtuelle Fachhochschule* (*VFH*) zur Verfügung gestellter anonymisierter Datenbestand aus dem Wintersemester 2021/2022[9]. Hierin enthalten sind die Moodle-Nutzungsdaten von Studenten, Dozenten und anderem Personal der *Berliner Hochschule für Technik* (*BHT*) und der *Alice Salomon Hochschule Berlin* (*ASH*) aus den folgenden Studiengängen:

- Master-Studiengang Medieninformatik Online (MMIO)
- Bachelor-Studiengang Wirtschaftsingenieurwesen Online (BWIO)
- Bachelor-Studiengang Wirtschaftsinformatik Online (BWINF)
- Bachelor-Studiengang Soziale Arbeit Online (BSAO)

2.3.1 Beschreibung der Daten

Um den Zugriff auf die Daten und deren praktische Untersuchung zu erleichtern, wurden diese zunächst seitens der VFH aus der Datenbank des Moodle-Systems (Green, 2022) extrahiert und in einem ersten Arbeitsschritt in nur einer Relation zusammengeführt.

Hierbei wurden Merkmale, die für diese Arbeit erwartungsgemäß keinen Mehrwert besitzen bereits eliminiert, während z. B. das Attribut *Studiengang* als neue Spalte in die Tabelle aufgenommen wurde, um die Zuordnung der Datensätze zu den jeweiligen Studiengängen unmittelbar erkennen zu können. Daneben wurden vorab die Merkmale *relateduserid*, *course_module_type* und *instanceid* eingefügt, um bei der Datenanalyse auch deren Informationsgehalt zur Identifikation typischen Benutzerverhaltens sinnvoll nutzen zu können.

Damit die Daten in einem beliebigen IT-Umfeld einfach weiterverarbeitet werden können, wurden sie im Anschluss an ihre Vorbereitung in einem für diesen Zweck

[9] Das gesamte Semester musste nach der SARS-CoV-2-Infektionsschutzmaßnahmenverordnung des Berliner Senates unter erhöhten Sicherheitsbedingungen stattfinden. Die Regelungen für das Lehr- und Prüfungsgeschehen wurden an der BHT infolgedessen wie folgt angepasst:
  - keine Lehrveranstaltungen und Prüfungen in Präsenz
  - keine Zählung des Semesters als Fachsemester
  - keine Zählung von Prüfungsfehlversuchen

typischen CSV-Format exportiert. Übergeben wurden die CSV-Daten schließlich als offene und komprimierte Textdateien in ASCII-Kodierung (Cerf, 1969), in der die Daten entgegen der üblichen Praxis jedoch nicht durch Kommata, sondern durch Semikola strukturiert waren.

Die freie Wahl eines Trennzeichens ist beim CSV-Format möglich, weil dieses nur allgemein beschreibt, wie die Tupel einer Relation und darin enthaltene Werte in der Regel interpretiert werden. Das Format definiert aber keinen verbindlichen Standard (Shafranovich, 2005), so dass die Daten entgegen ihrer Definition als Comma-Separated Values nicht zwingend nur durch Kommata zu strukturieren sind.

Der zur Verfügung gestellte Datenbestand umfasst insgesamt 969032 Datensätze. Dabei handelt es sich genau betrachet um eine spezifische Teilmenge von Loggings auf dem Moodle-Server der VFH, mit denen client- und serverseitige Aktionen fortlaufend protokolliert werden. Typische Aktionen, die so u. a. aufgezeichnet werden sind das Aufrufen eines Kursmoduls, das Starten eines Uploads, das Senden einer Nachricht oder auch die Bewertung einer Aufgabe.

Formale Angaben über die Daten

Erste interessante Einblicke in die Art, den Umfang und die Struktur der zu untersuchenden Daten ergeben sich nach deren Import in eine MySQL-Datenbank durch einfache statistische SQL-Abfragen:

Abfrage zu Art und Umfang der implementierten MySQL-Datenbank

```
mysql> SELECT TABLE_SCHEMA, TABLE_NAME, ENGINE,
      (SELECT COUNT(*) FROM moodle_data) AS TABLE_ROWS, TABLE_COLLATION
       FROM information_schema.tables WHERE table_name = "moodle_data";
+-----------------+-------------+--------+------------+------------------+
| table_schema    | table_name  | engine | table_rows | table_collation  |
+-----------------+-------------+--------+------------+------------------+
| vfh_moodle_ws21 | moodle_data | InnoDB |     969032 | ascii_general_ci |
+-----------------+-------------+--------+------------+------------------+
```

Die Ergebnistabelle zeigt einen Ausschnitt der Metadaten, die standardmäßig vom MySQL-Server in der Datenbank INFORMATION_SCHEMA zu jeder verwalteten

Datenbank gespeichert werden.[10] Die aufgelisteten Werte informieren u. a. über das Speichersystem *InnoDB*, die Anzahl der Datensätze *969032* und die Kollation, die definiert, nach welchen Regeln Zeichenketten verglichen werden. Übereinstimmend mit der ASCII-Kodierung der CSV-Daten wurde bei der Erstellung der Datenbank *vfh_moodle_ws21* die Kollation *ascii_general_ci* gewählt.

Abfrage zu Struktur und Inhalt der importierten Originaldaten

```
mysql> DESCRIBE moodle_data;
+--------------------+-------------+------+-----+---------+-------+
| Field              | Type        | Null | Key | Default | Extra |
+--------------------+-------------+------+-----+---------+-------+
| courseid           | int(11)     | YES  |     | NULL    |       |
| Studiengang        | varchar(11) | YES  |     | NULL    |       |
| userid             | int(11)     | YES  | MUL | NULL    |       |
| relateduserid      | int(11)     | YES  |     | NULL    |       |
| action             | varchar(10) | YES  |     | NULL    |       |
| eventname          | varchar(57) | YES  |     | NULL    |       |
| objecttable        | varchar(27) | YES  |     | NULL    |       |
| objectid           | int(11)     | YES  |     | NULL    |       |
| timecreated        | int(11)     | YES  |     | NULL    |       |
| course_module_type | varchar(18) | YES  |     | NULL    |       |
| instanceid         | int(11)     | YES  |     | NULL    |       |
+--------------------+-------------+------+-----+---------+-------+
```

Die obige Ausgabe beschreibt das Schema der importierten Daten. Von Interesse für diese Arbeit sind hier aber nur die Werte zu *Field* und *Type*, die die Spaltennamen der Tabelle und die Datentypen der darin enthaltenen Werte angeben.

Informationen und deren Beziehungen

Die nachfolgende tabellarische Übersicht zeigt nun, welche Informationen in den Feldern der verschiedenen Merkmale des Datenbestandes tatsächlich enthalten sind und in welchen Beziehungen diese innerhalb der aktuell betriebenen relationalen Datenbank des VFH-Moodle stehen.[11]

[10] Siehe auch die MySQL Documentation: 24.1 Introduction, MySQL 5.7 Reference Manual, 05/2022
[11] Siehe auch die Moodle Entity Relationship Documentation (Green, 2022): Moodle ERD, 05/2022

Merkmal	**Information / *Beziehung innerhalb des VFH-Moodle***
courseid	Studienmodul, das im WS 2021/2022 belegt wurde. *Fremdschlüssel zur Identifikation eines bestimmten Studienmoduls in der Relation course.*
Studiengang	Studiengang, in dem aktuell studiert wird. *Frei gewählte Kennziffer zur eindeutigen Unterscheidung der Studiengänge; bedeutet keine Referenz auf eine andere Entität.*
userid	Kennzahl zur Identifikation des Benutzers. *Fremdschlüssel zur Identifikation eines bestimmten Benutzers (z. B. der Sender einer Nachricht) in einer von der VFH für diesen Zweck bereitgestellten weiteren Relation.*
relateduserid	Kennzahl zur Identifikation eines weiteren Benutzers. *Fremdschlüssel des interagierenden Benutzers, der z. B. bei einem Chat den Empfänger einer Nachricht repräsentiert.*
action	Interaktion, die im Moodle-System ausgeführt wurde. *Allgemeinere Form des eventtype, der auch im eventname als notwendiger Bestandteil redundant enthalten ist.*
eventname	Mehrteiliger Bezeichner für das ausgelöste Event. *Ausgelöst duch eine Interaktion wird ein Bezeichner durch die drei Werte modulename, instance und eventtype der Relation event generiert und eingetragen.*
objecttable	Relation zur Verwaltung von Objekttabellen. *Abhängig von der Art des Kursmoduls und der Interaktion werden die durch Verwendung bestimmter Objekte tangierten Tabellen dokumentiert, z. B. assign_grades, course_modules oder forum_discussions*
objectid	Kennzahl zur Identifikation des verwendeten Objekts. *Fremdschlüssel zur Identifikation des durch die Interaktion tangierten Objekts in der zugehörigen Relation objecttable.*
timecreated	Zeitpunkt der ausgeführten Interaktion. *10-stelliger Unix Epoch Timestamp, der seit Donnerstag, dem 01.01.1970, 00:00 Uhr UTC die vergangenen Sekunden zählt.*
course_module_type	Typ des verwendeten Kursmoduls. *Zur Anreicherung des Informationsgehalts aus der Relation course_modules entnommener Bezeichner des Modultyps, z. B. assign, forum, label oder resource*
instanceid	Kennzahl zur Identifikation des Kursmodultyps. *Fremdschlüssel zur Identifikation des Kursmodultyps in der zugehörigen Relation course_modules.*

Tabelle 1: Schema des Datenbestandes mit Erläuterungen

2.3.2 Vorbereitung der Daten

Die Datenqualität spielt bei der Datenanalyse eine sehr große Rolle. Daten müssen zwingend in einer entsprechend hohen Qualität vorliegen, damit im Nachhinein die gewonnenen Analyseergebnisse als fundiert gelten dürfen.

Wichtige Kriterien der Datenqualität sind u. a. die Vollständigkeit, die Richtigkeit sowie die Eindeutigkeit der Daten (Wang & Strong, 1996). Daneben ist aber auch die eigentliche Relevanz von grundlegender Bedeutung, da die Einbeziehung nicht relevanter Daten in eine Untersuchung die daraus resultierenden Ergebnisse stark verfälschen kann.

Mit Blick auf den Untersuchungsgegenstand dieser Arbeit – *das studentische Lern- und Kommunikationsverhalten* – wurde folglich entschieden, jene Datensätze die sich nicht sicher auf Aktivitäten von Studenten beziehen, zu kennzeichnen und bei den anschließenden Untersuchungen zu ignorieren.

Die praktische Unterscheidung von Studenten und anderen Benutzern, wie z. B. Dozenten, Studiengangskoordinatoren oder weiterem Personal der Hochschulen, erfolgte anschließend in mehreren Schritten:

Abfrage zur Feststellung der Anzahl aller Benutzer

```
mysql> SELECT COUNT(DISTINCT userid) AS "total_number_users"
       FROM moodle_data
       WHERE !(userid = -2 OR userid = -3);
+--------------------+
| total_number_users |
+--------------------+
|                142 |
+--------------------+
```

In der SQL-Abfrage berücksichtigt und im Ergebnis damit bereits exkludiert sind zwei Benutzergruppen, die einer Beobachtung ihres Verhaltens nicht zugestimmt haben (userid = -2) oder im Bachelor-Studiengang Medieninformatik aktiv waren (userid = -3). Abzüglich dieser beiden Personengruppen ergeben sich mithin die obigen 142 Einzelpersonen.

Studenten und andere Benutzer unterscheiden sich durch gewisse Interaktionen, die auf rollenspezifischen Berechtigungen beruhen und im Idealfall nur von einer

der beiden Gruppen ausgeführt werden können: So wurde für den konkreten Fall angenommen, dass Studenten generell keine Berechtigung zur Notenvergabe haben und hierüber entsprechend identifiziert werden können.

Abfragen zur Identifikation von Studenten

```
mysql> SELECT COUNT(DISTINCT userid) AS "total_number_students"
       FROM moodle_data
       WHERE userid
            NOT IN(SELECT userid FROM moodle_data WHERE action = "graded")
            AND !(userid = -2 OR userid = -3);
+-----------------------+
| total_number_students |
+-----------------------+
|                    69 |
+-----------------------+
```

Die obige Abfrage ist so konzipiert, dass sie im Unterschied zu der vorhergehenden auch noch diejenigen Benutzer exkludiert, denen für das Merkmal *action* der Wert *graded* zugeordnet ist. Damit werden von den vorher ermittelten 142 Einzelpersonen all jene Benutzer abgezogen, die an Notenvergaben aktiv beteiligt waren.

Die nachfolgende Abfrage gibt nun eine Liste der ermittelten 69 Studenten aus, auf eine Darstellung der Ergebnistabelle an dieser Stelle soll aber aus Platzgründen verzichtet werden.

```
mysql> SELECT userid AS "students"
       FROM moodle_data
       WHERE userid
            NOT IN(SELECT userid FROM moodle_data WHERE action = "graded")
            AND !(userid = -2 OR userid = -3)
       GROUP BY userid
       ORDER BY userid;
```

Ergänzend sei hier erwähnt, dass mit den Ergebnissen aus den obigen Abfragen zur Identifikation die von der Hochschule genannte Menge von 75 Studenten nur angenähert werden, und eventuell werden dadurch auch studentische Mitarbeiter ausgeklammert, die bei Bewertungen mitgewirkt haben. Andere Einflussfaktoren wie z. B. spezifische *eventtypes* bringen hier zwar ein wenig Verbesserung, erhöhen aber auch deutlich das Fehlerrisiko und blieben infolgedessen unbeachtet.

Um im weiteren Verlauf der Arbeit die Auswahl der identifizierten Studenten zu vereinfachen und damit auch den zu analysierenden Datenbestand zu reduzieren, wurde entschieden eine Umbenennung all jener Benutzer durchzuführen, die nicht sicher als Studenten gelten können.

Hierzu wurde zunächst in der ebenfalls von der Hochschule zur Verfügung gestellten (Hilfs-)Relation zur Identifikation der (anonymisierten) Benutzerkennung für das *Hochschulpersonal* ein Datensatz mit dem Primärschlüssel *-1* eingefügt.

Einfügen eines neuen Datensatzes in die Relation moodle_data_users

```
mysql> INSERT INTO moodle_data_users
       VALUES (-1, "Hochschulpersonal");
```

Anschließend wurden alle Datensätze geändert, die für das Merkmal *userid* einen Wert besitzen, der in mindestens einem anderen Datensatz der Tabelle auch mit dem Wert *graded* für das Merkmal *action* in Beziehung steht. In all diesen Fällen wurde die bestehende *userid* mit dem Wert *-1* überschrieben und damit eine Referenz auf den Primärschlüssel in der Hilfsrelation *moodle_data_users* hergestellt.

Änderung der userid für das Hochschulpersonal

```
mysql> UPDATE moodle_data
       SET userid = -1
       WHERE userid
            IN (SELECT *
                FROM(SELECT md1.userid
                     FROM moodle_data md1
                     INNER JOIN moodle_data md2
                     ON md1.userid = md2.userid
                     WHERE md2.action = "graded")
                AS moodle_data_temp);
```

Folgende vereinfachte Abfragen ergeben nun genau die oben bereits nach Art und Umfang ermittelten Studenten und bestätigen so deren erfolgreiche Identifikation.

Abfragen zur Identifikation von Studenten

```
mysql> SELECT userid AS "students"
       FROM moodle_data
       WHERE userid > 0
       GROUP BY userid
       ORDER BY userid;
```

```
mysql> SELECT COUNT(DISTINCT userid) AS "total_number_students"
       FROM moodle_data
       WHERE userid > 0;
+-----------------------+
| total_number_students |
+-----------------------+
|                    69 |
+-----------------------+
```

Abschließend sei noch angemerkt, dass mit dieser Änderung in den Datensätzen in denen zuvor die *userid* auf *-2* gesetzt war, diese mit dem Wert *-1* überschrieben wurde. Dies ist nachvollziehbar und richtig, da die Mitarbeiter der Hochschule nicht Gegenstand der Untersuchung sind, und demnach generell keine Zustimmung zur Beobachtung ihres Verhaltens geben dürfen.

Erste Einblicke in die Arbeitsdaten

Nach Vorbereitung der Daten in den vorigen Abschnitten soll im Folgenden anhand einiger statistischer Abfragen der tatsächliche Gegenstand der Untersuchung (Arbeitsdaten) genauer betrachtet und mithin erste Erkenntnisse daraus gewonnen werden.

Abfrage zum Umfang des zu untersuchenden Datenbestands

```
mysql> SELECT COUNT(*) AS 'table_rows_students'
       FROM moodle_data
       WHERE userid > 0;
+---------------------+
| table_rows_students |
+---------------------+
|              234664 |
+---------------------+
```

Wie das Ergebnis zeigt, hat sich die Anzahl der für die Untersuchung relevanten Datensätze deutlich verringert. Von den ursprünglich 969032 Datensätzen sind nach Vorbereitung der Daten nur noch 234664 verblieben (entspricht einer Reduktion um etwas mehr als 75 Prozent).

Abfrage zur Anzahl der Datensätze pro Student

```
mysql> SELECT userid, COUNT(userid) AS "total_number_records"
       FROM moodle_data
       WHERE userid > 0
       GROUP BY userid;
+--------+----------------------+
| userid | total_number_records |
+--------+----------------------+
|      1 |                 3865 |
|      2 |                 4706 |
|      3 |                 3373 |
     ...                    ...
|     21 |                24595 |
     ...                    ...
|    142 |                   10 |
|    143 |                 1387 |
|    144 |                  240 |
+--------+----------------------+
69 rows in set (0,05 sec)
```

Aus Platzgründen werden in der obigen Ergebnistabelle nur wenige der insgesamt 69 Zeilen des Abfrageergebnisses angezeigt. Es wird aber auch bereits in diesem kleinen Ausschnitt deutlich, wie unterschiedlich die studentischen Aktivitäten über das Semester hinweg in ihrem Umfang waren.

Abfrage zur Anzahl der Studenten pro Studiengang

```
mysql> SELECT Studiengang, COUNT(DISTINCT userid) AS "total_number_students"
       FROM moodle_data
       WHERE userid > 0
       GROUP BY Studiengang;
+-------------+-----------------------+
| Studiengang | total_number_students |
+-------------+-----------------------+
| 0           |                    69 |
| 1           |                     8 |
| 2           |                    20 |
| 3           |                    26 |
| 4           |                    12 |
+-------------+-----------------------+
```

In der Ausgabe enthalten ist neben den oben genannten Studiengängen 1 bis 4 überraschenderweise ein weiterer Studiengang 0. Hierbei handelt es sich jedoch nicht

um einen Studiengang wie die anderen vier, sondern um eine spezifische Entität, die sich auf die Aktivitäten bezieht, die außerhalb des eigentlichen Kursgeschehens stattfinden, z. B. Logins, Chats oder Aufrufe des Kalenders bzw. Dashboards.

Bemerkenswert ist hier auch, dass dem Studiengang 0 alle zuvor identifizierten Studenten zugeordnet sind, die Anzahl der Studenten in den Studiengängen 1 bis 4 dagegen nur 66 beträgt. Unter Berücksichtigung dessen, dass die Studiengänge von der Hochschule anhand der darin enthaltenen Kurse abgeleitet wurden, ließe sich so u. a. folgern, dass selbst bei Aktivitäten in mehreren Studiengängen mindestens drei Studenten in diesem Semester keinen normalen Kurs im Sinne eines Pflicht- oder Wahlpflichtmoduls besucht haben.

Abfrage zur Anzahl der Kursmodule pro Student

```
mysql> SELECT userid, COUNT(DISTINCT courseid) AS "total_number_courses"
       FROM moodle_data
       WHERE Studiengang > 0 AND userid > 0
       GROUP BY userid
       ORDER BY total_number_courses;
+--------+----------------------+
| userid | total_number_courses |
+--------+----------------------+
|     60 |                    1 |
     ...                    ...
|     84 |                    6 |
     ...                    ...
|    108 |                   12 |
     ...                    ...
|     29 |                   18 |
|    118 |                   21 |
|    133 |                   22 |
+--------+----------------------+
58 rows in set (1,91 sec)
```

Auch wenn die Tabelle die Ergebnisse aus Platzgründen wiederum nur teilweise darstellt, so lassen sich zwei weitere interessante Erkenntnisse schnell ableiten:

- Elf Studenten haben in den Studiengängen 1 bis 4 keinen Kurs besucht.
- Die Menge an Kursen pro Student lag teilweise weit über der empfohlenen Anzahl von 6 Kursmodulen für ein Vollzeitstudium in Regelstudienzeit.

Mit Blick auf die unerwartet hohen Anzahlen an Kursen pro Student soll zum Schluss dieses Kapitels die Menge an Studenten mit überdurchschnittlich vielen Kursen und das Verhältnis der Studiengänge pro Student untersucht werden.

Abfrage zu Studenten mit überdurchschnittlich vielen Kursen

```
mysql> SELECT userid, COUNT(DISTINCT courseid) AS "total_number_courses"
       FROM moodle_data
       WHERE Studiengang > 0 AND userid > 0
       GROUP BY userid
       HAVING total_number_courses >= 12
       ORDER BY total_number_courses;
+--------+----------------------+
| userid | total_number_courses |
+--------+----------------------+
|    109 |                   12 |
     ...                     ...
|    108 |                   12 |
|    122 |                   13 |
|     91 |                   13 |
|     24 |                   14 |
|     18 |                   17 |
|     29 |                   18 |
|    118 |                   21 |
|    133 |                   22 |
+--------+----------------------+
18 rows in set (1,77 sec)
```

Abfrage zur Anzahl der Studiengänge pro Student

```
mysql> SELECT userid, COUNT(DISTINCT Studiengang) AS "total_number_studies"
       FROM moodle_data
       WHERE Studiengang > 0 AND userid > 0
       GROUP BY userid
       HAVING total_number_studies > 1
       ORDER BY total_number_studies;
+--------+----------------------+
| userid | total_number_studies |
+--------+----------------------+
|     81 |                    2 |
     ...                     ...
|     55 |                    2 |
|     21 |                    3 |
+--------+----------------------+
7 rows in set (1,90 sec)
```

Auch die letzten beiden Abfragen können mit ihren Ergebnissen überraschen. So waren 18 von 58 Studenten (31%) über das Semester hinweg in mindestens doppelt so vielen Kursen aktiv, wie es von den Hochschulen für ein Vollzeitstudium in der Regel empfohlen wird.

Der Gedanke, dass es dann auch Studenten gegeben haben könnte, die vielleicht sogar mehr als einen Studienang besucht haben, wird durch die Abfrage zur Anzahl der Studiengänge pro Student eindrucksvoll bestätigt: Insgesamt sieben Studenten waren in mehr als einem der eingangs genannten Studiengänge 1 bis 4 tätig.

2.3.3 Visualisierung der Daten

...

3 Analyse

...

4 Ergebnisse

...

5 Fazit

...

6 Ausblick

...

Literaturverzeichnis

Azevedo, A. & Santos, M. (2008, 01). KDD, SEMMA and CRISP-DM: A parallel overview. In (S. 182-185).

Cerf, V. (1969, Oktober). *ASCII format for network interchange* (Nr. 20). RFC 20. RFC Editor. Zugriff auf https://www.rfc-editor.org/info/rfc20 doi: 10.17487/RFC0020

Fayyad, U., Piatetsky-Shapiro, G. & Smyth, P. (1996, Mar.). From data mining to knowledge discovery in databases. *AI Magazine*, *17* (3), 37. Zugriff auf https://ojs.aaai.org/index.php/aimagazine/article/view/1230 doi: 10.1609/aimag.v17i3.1230

Green, M. (2022). *The Moodle Database. Table and relationship documentation generated from moodle source code.* Zugriff am 2022-04-08 auf https://www.examulator.com/er/

Runkler, T. A. (2020). Introduction. In *Data analytics: Models and algorithms for intelligent data analysis* (S. 1–4). Wiesbaden: Springer Fachmedien. Zugriff auf https://doi.org/10.1007/978-3-658-29779-4_1 doi: 10.1007/978-3-658-29779-4_1

Shafranovich, Y. (2005, Oktober). *Common Format and MIME Type for Comma-Separated Values (CSV) Files* (Nr. 4180). RFC 4180. RFC Editor. Zugriff auf https://www.rfc-editor.org/info/rfc4180 doi: 10.17487/RFC4180

Shearer, C. (2000). The crisp-dm model: The new blueprint for data mining. *Journal of Data Warehousing*, *5* (4).

Wang, R. Y. & Strong, D. M. (1996). Beyond accuracy: What data quality means to data consumers. *J. Manag. Inf. Syst.*, *12*, 5-33.

Erklärung zur Urheberschaft

Ich habe die Arbeit selbständig verfasst, keine anderen als die angegebenen Quellen und Hilfsmittel benutzt, sowie alle Zitate und Übernahmen von fremden Aussagen kenntlich gemacht.

Die Arbeit wurde bisher keiner anderen Prüfungsbehörde vorgelegt.

Die vorgelegten Druckexemplare und die vorgelegte digitale Version dieser Arbeit sind vollkommen identisch.

Heidelberg, dd.mm.2022

_____

Unterschrift

Inhalt des beigefügten Datenträgers

Verzeichnis / Beschreibung
`/1_` ...
`/2_` ...
`/3_` ...