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 insbesondere zur angewandten Methodik und zu dem zu untersuchenden Gegenstand in dieser Arbeit 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 und die im Zuge dessen eingesetzten technischen Ressourcen vorgestellt.

Unter verschiedenen Aspekten wird im Anschluss 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.

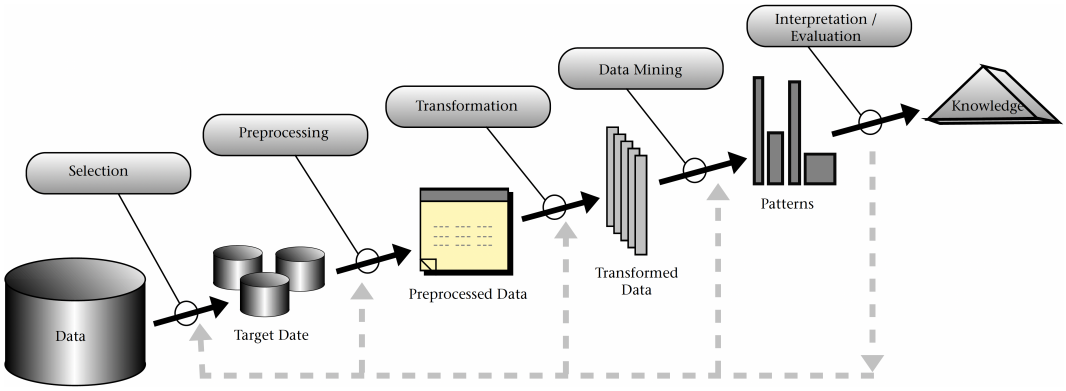

Abbildung 1: Phasen des KDD-Prozesses. Original von Fayyad et al. (1996).

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).

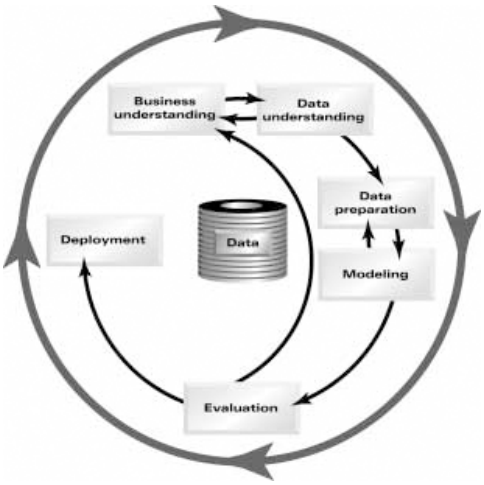

Abbildung 2: Phasen des CRISP-DM. Original von Shearer (2000).

[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 vorherigen Abschnitt präsentierten Vorgehensmodelle haben alle dasselbe Ziel: Sie möchten 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

Für die anstehenden Untersuchungen im Rahmen dieser Arbeit soll, wie eingangs beschrieben, nun auf Basis des *CRISP-DM* das im weiteren Verlauf verwendete Vorgehensmodell in Grundzügen skizziert werden. Dieses wird später bei der tatsächlichen Durchführung der Datenanalyse dann erneut als Vorlage dienen und je nach Anforderung noch einmal spezifisch angepasst werden müssen.

Hier kommen Erläuterungen zu den für diese Arbeit relevanten und auch nicht relevanten Prozessphasen –> Darstellung des eigenen (vereinfachten) Ablaufs zur Durchführung – standardisierter – Untersuchungen und damit einhergehender Vergleichbarkeit der Ergebnisse. Hinweise auf die Teilabschnitte (1. Datenaufbereitung, ...) des Verfahrens.
Diagramm zum eigenen Vorgehensmodell erstellen (s. Exposé oder Runkler).

2.1.3 Explorative Datenanalyse

...

2.1.4 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[6]. In diesem enthalten sind die Moodle-Nutzungsdaten von Studenten der *Berliner Hochschule für Technik* (*BHT*) sowie der *Alice Salomon Hochschule Berlin* (*ASH*) aus den folgenden vier Online-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 Vorbereitung der Daten

Um die Übersichtlichkeit der Daten und deren Untersuchung im Rahmen dieser Arbeit zu erleichtern, wurden diese in einem ersten Schritt aus der umfangreichen Datenbank des VFH-Moodle (Green, 2022) extrahiert und in einer einzigen Relation zusammengefasst.

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 verschiedenen Datensätze zu den jeweiligen Studiengängen unmittelbar erkennen zu können. Des weiteren wurden vorab die beiden Merkmale *course_module_type* und *instanceid* eingefügt, um auch deren Informationsgehalt sinnvoll nutzen zu können.

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. Kriterien der Daten-

[6] 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

qualität sind dabei u. a. die Vollständigkeit, die Richtigkeit und auch die Eindeutigkeit der Daten (Wang & Strong, 1996). Daneben ist aber auch deren eigentliche Relevanz von grundlegender Bedeutung, da die Einbeziehung nicht relevanter Daten in eine Untersuchung die daraus resultierenden Ergebnisse stark negativ beeinflussen kann.

Mit Blick auf den Untersuchungsgegenstand dieser Arbeit – *das studentische Lern- und Kommunikationsverhalten* – wurde bezüglich der vorliegenden Daten daher einvernehmlich entschieden, alle Datensätze die sich auf Aktivitäten der Dozentenschaft beziehen, im Vorfeld zu entfernen.

TEXTALTERNATIVE 1: *Bei der notwendigen Identifikation der Dozenten wurde schließlich auf deren besondere Berechtigungen zu bestimmten Interaktionen abgestellt. Die gemeinsame Betrachtung der Werte für die Merkmale userid und eventname hat diese Annahme bestätigt und gezeigt (s. nachfolgender Scatterplot), dass über die Gesamtdauer des Semesters nur eine gewisse Gruppe von 39 Benutzern ein Event vom Typ course_module_created ausgelöst hat. Im Gegensatz dazu hat eine Gruppe von 70 Benutzern über die gleiche Zeit ein Event vom Typ response_submitted initiiert. Diese typischen Aktivitäten und die klare disjunkte Aufteilung dieser beiden Gruppen, die gemeinsam wiederum die Gesamtanzahl von 109 Personen repräsentieren, wurde als klares Indiz für eine Dozentätigkeit der ersten Gruppe gewertet und mithin konnten im Anschluss die für die weiteren Untersuchungen irrelevanten Datensätze aus dem Datenbestand gelöscht werden.*

TEXTALTERNATIVE 2: *Bei der nachfolgenden Identifikation der Dozenten wurde schließlich festgestellt, dass manche userids mit mehr als nur einem Studiengang in Beziehung stehen. Dies wurde als klares Indiz für eine Dozentätigkeit gewertet und mithin konnten im Anschluss die für die weiteren Untersuchungen irrelevanten Datensätze aus dem Datenbestand gelöscht werden.*

2.3.2 Beschreibung der Daten

Die allgemeine Beschreibung der Daten und die Darstellung der statistischen Informationen auf Hinweis von Frau Dr. Sauer besser in getrennten Unterkapiteln unterbringen.

Damit die Daten in einem beliebigen IT-Umfeld einfach weiterverarbeitet werden

können, wurden sie im Anschluss an ihre Vorbereitung in einem für diesen Zweck 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 288.152 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.

Hier eventuell noch einen Abschnitt zu den Skalenniveaus der jeweiligen Merkmale ergänzen: Es gibt bis auf den Zeitstempel nur qualitative nominalskalierte Daten (kategoriale Daten). Damit sind lediglich Operationen wie die Überprüfung auf Gleichheit und Ungleichheit zulässig oder die Bildung des Modus, auch Modalwert genannt, also dem am häufigst vorkommenden Wert. (s. Kapitel 2. Daten). Der Zeitstempel selbst ist ein quantitatives proportionalskaliertes und diskretes Merkmal, das dagegen alle Grundrechenarten erlaubt? Noch einmal überpüfen!

Die folgende tabellarische Übersicht zeigt, welche Informationen in den Feldern der verschiedenen Merkmale des Datenbestandes enthalten sind und in welchen Beziehungen diese innerhalb der relationalen Datenbank des VFH-Moodle stehen; siehe hierzu auch die Moodle Entity Relationship Documentation (Moodle-ERD) (Green, 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. *Hash-Code zur Anonymisierung der Benutzerkennung, über die sonst ein Benutzer in der Relation user konkret referenziert werden kann.*
action	Interaktion, die im Moodle-System ausgeführt wurde. *Eindeutiger Wert (Fremdschlüssel?), der aus einer Relation außerhalb des eigentlichen VFH-Moodle bezogen wird.*
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 vergangene Sekunden seit Donnerstag, dem 01.01.1970, 00:00 Uhr UTC 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

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

1. Abfrage: 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 |     288152 | ascii_general_ci |
+-----------------+-------------+--------+------------+------------------+
```

Zu den jeweiligen Abfrageergebnissen immer auch noch einen erläuternden Text ergänzen, so z.B. zum Typ der engine (hier InnoDB) oder der table_collation ... so i.S. von 'In der Ergebnistabelle ist ersichtlich ...'

2. Abfrage: Datentypen und Constraints der importierten CSV-Daten

```
mysql> DESCRIBE moodle_data;
+--------------------+-------------+------+-----+---------+-------+
| Field              | Type        | Null | Key | Default | Extra |
+--------------------+-------------+------+-----+---------+-------+
| courseid           | int(11)     | YES  |     | NULL    |       |
| Studiengang        | varchar(11) | YES  |     | NULL    |       |
| userid             | varchar(32) | 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    |       |
+--------------------+-------------+------+-----+---------+-------+
```

3. Abfrage: Summe aller Datensätze pro Studiengang

```
mysql> SELECT Studiengang, COUNT(Studiengang) AS "total number records"
       FROM moodle_data
       GROUP BY Studiengang;
+-------------+----------------------+
| Studiengang | total number records |
```

```
+-------------+----------------------+
|             |                12051 |
| 1           |                14115 |
| 2           |                53706 |
| 3           |               111163 |
| 4           |                97117 |
+-------------+----------------------+
```

4. Abfrage: Summe aller Kurse in den Studiengängen 1 bis 4

```
mysql> SELECT COUNT(DISTINCT courseid) AS "total number courses"
       FROM moodle_data
       WHERE Studiengang != "";
+----------------------+
| total number courses |
+----------------------+
|                  108 |
+----------------------+
```

5. Abfrage: Summe aller Benutzer in den Studiengängen 1 bis 4

```
mysql> SELECT COUNT(DISTINCT userid) AS "total number users"
       FROM moodle_data
       WHERE Studiengang != "";
+--------------------+
| total number users |
+--------------------+
|                107 |
+--------------------+
```

2.3.3 Visualisierung der Daten

Visualisierungen sollen helfen, ein besseres Verständnis über die Datenbasis zu bekommen.

3 Umsetzung

...

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_` ...