Kriterien für die automatisierte Bewertung von user-generated educational Microcontent

Oliver Ott[1], Michael Hielscher[2]

[1]Institut für Medienbildung, Pädagogische Hochschule Bern
Helvetiaplatz 2, CH-3005 Bern
oliver.ott@phbern.ch

[2]Institut für Medien und Schule, Pädagogische Hochschule Schwyz
Zaystrasse 42, CH-6410 Goldau
michael.hielscher@phsz.ch

Abstract: Im World Wide Web entstanden in den letzten Jahren diverse Austausch-Plattformen mit freien Lern- und Lehrinhalten (OER). Bei Web-2.0-Diensten wie z.B. GeoGebra.org, Educaplay.com oder LearningApps.org stellen Autorinnen und Autoren große Mengen von nutzergenerierten Inhalten bereit, die qualitativ sehr unterschiedlich sind. Es stellt sich die Frage, ob diese Inhalte anhand von statistischen Kriterien automatisiert bewertet werden können. Der Artikel beschreibt einige mögliche und praktisch umsetzbare Bewertungskriterien, die exemplarisch an der Plattform LearningApps.org untersucht wurden.

1 Einleitung

Seit den 1990er Jahren wurden unzählige Bildungsplattformen und Bildungsserver aufgebaut, die sich zum Ziel gesetzt haben, kostenlose digitale Lehr- und Lernmaterialien zusammenzutragen und zentral zur Verfügung zu stellen. Die Bandbreite reicht dabei von streng redaktionell betreuten Portalen bis hin zu offenen Austausch-Plattformen ohne jegliche Qualitätskontrolle. Seit der Entstehung des Web 2.0 tragen Nutzerinnen und Nutzer zunehmend selbst Inhalte zu diesen Plattformen bei. Aktuell spricht man in diesem Zusammenhang häufig auch von Open Educational Resources (OER). Als zentrale Merkmale von OER nennt Geser [Ge07] einen freien und kostenlosen Zugang und eine Lizenz, welche die Weiterverarbeitung und Rekombination der Materialien erlaubt. Zudem werde zur Nutzung bzw. Herstellung der Inhalte offene und freie Software (Open-Source) eingesetzt.

Bei Web-2.0-Plattformen für Bildungsinhalte nach dem Vorbild von YouTube und Flickr werden täglich hunderte neue Inhalte veröffentlicht. Durch das große Angebot steigt theoretisch auch die Chance, einen passenden Beitrag für den eigenen Unterricht zu finden. Die Selektion und Pflege der Inhalte wird jedoch den Nutzenden überlassen, was häufig zu einer sehr heterogenen Qualität führt. Dies führt zum Problem der aufwändigen Suche und Filterung von brauchbaren und weniger brauchbaren Inhalten. Typischerweise erlauben Web-2.0-Plattformen ihren Nutzenden eingestellte Inhalte etwa mit Sternen zu bewerten, um wiederum anderen Nutzenden die Suche zu erleichtern.

Bewertungen sind ein schnelles und in der Regel anonymes Feedback für die Autorinnen und Autoren, werden jedoch eher selten abgegeben. Laut Siersdorfer et al. [Si10] wird bei YouTube im Durchschnitt nur bei 0,22% der Aufrufe eine Bewertung abgegeben oder ein Kommentar hinterlassen.

Die Rangierung und Filterung von Inhalten anhand der Nutzer-Bewertungen kann auf Bildungsplattformen mit deutlich weniger Aufrufen nur bedingt funktionieren, da schlicht zu wenige Daten zur Verfügung stehen. Dennoch bieten die meisten Bildungsplattformen diese Sortierung für ihre Nutzenden an. In dieser Arbeit werden Kriterien für eine automatisierte Bewertung von Inhalten als Grundlage für einen verbesserten Sortieralgorithmus vorgeschlagen. Die Kriterien wurden anhand der statistischen Daten der Plattform LearningApps.org [HHR13] [Hi13] evaluiert.

2 Web-2.0-Dienste für user-generated educational Microcontent

Im Schulumfeld gibt es eine Vielzahl von Web-2.0-Diensten, die Lernbausteine bzw. Materialien für kurze Lernsequenzen sammeln und als user-generated educational Microcontent anderen Nutzenden zur Verfügung stellen. Unter Lernbausteinen sollen hier interaktive Lernumgebungen, Übungs- und Anwendungsaufgaben verstanden werden, die für Lernende häufig ein automatisiertes Feedback bieten und sich idealerweise in unterschiedlichen Lehr-/Lernszenarien wiederverwenden lassen. Diese Art von Inhalten wird beispielsweise von Plattformen wie GeoGebra.org, Educaplay.com und LearningApps.org angeboten.

GeoGebra.org ist eine beliebte Mathematiksoftware mit integrierten Visualisierungs- und Übungsmöglichkeiten. Nach dem Vorbild von YouTube bietet das Autorenwerkzeug seit 2011 GeoGebraTube an, wo GeoGebra-Applets ausgetauscht werden können. Das Angebot umfasste Anfang 2014 ca. 75'000 freie und interaktive Inhalte zu Geometrie, Algebra, Statistik und Analysis. Nutzende können Inhalte mit einem „Daumen hoch" bewerten und die Inhalte nach der Anzahl Bewertungen sortieren (Abbildung 1).

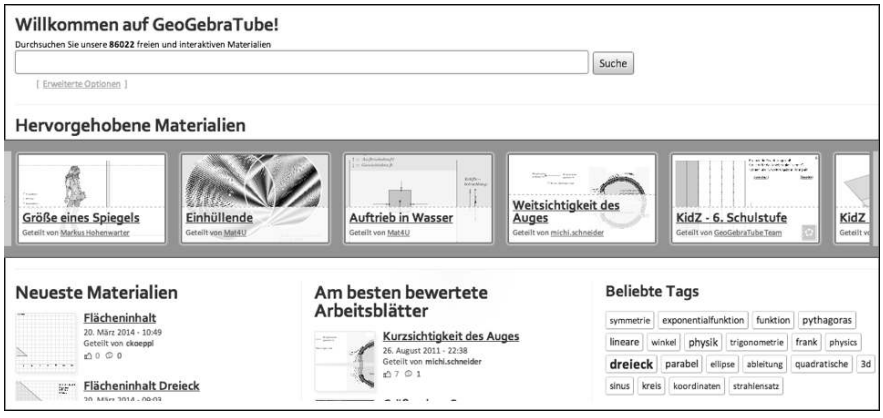

Abbildung 1: Web-2.0-Dienst GeoGebraTube

Andere Dienste stellen fächerunabhängige, webbasierte Autorenwerkzeuge für interaktive und multimediale Lernbausteine zur Verfügung. Auf Educaplay.com wurden Anfang 2014 rund 290'000 Lerninhalte in 15 Sprachen angeboten. LearningApps.org bietet ein ähnliches, noch stärker auf die Nutzung von multimedialen Inhalten zugeschnittenes Autorenwerkzeug an und umfasste Mitte 2014 ca. 390'000 Bausteine. Davon wurden rund 55'000 Inhalte von den Autorinnen und Autoren, in einem nach Fächern sortierten Katalog, zur freien Nutzung und Überarbeitung veröffentlicht. Auf Educaplay.com können Nutzende ähnlich wie bei GeoGebra.org über eine „Like"-Funktion Bewertungen vornehmen und die Inhalte nach der Anzahl Bewertungen sortieren. Bei LearningApps.org wurde eine differenziertere Bewertung mit 1-5 Sternen verwendet. Nutzende können die Inhalte nach absteigender Sterne-Bewertung sortieren (Abb. 2).

Abbildung 2: Web-2.0-Dienste Educaplay.com und LearningApps.org

3 Automatisierte Bewertung von user-generated educational Microcontent

Einfache Sortierprinzipien, etwa nach Anzahl Aufrufen oder neuesten Inhalten, sind zwar technisch leicht zu realisieren, bieten den Nutzenden jedoch wenig Hilfestellung bei der Auswahl von qualitativ guten und für sie relevanten Inhalten. Um die Inhalte in eine möglichst sinnvolle Reihenfolge zu bringen, müssen deshalb weitere Kriterien beigezogen und gewichtet werden. Dienste wie z.B. YouTube oder Flickr verwenden dazu eine Vielzahl von Daten über die Autorinnen und Autoren, die Nutzenden und deren Konsumverhalten. Recherchierende können bei Flickr die Bilder nach „Relevant" und „Interessant" filtern, wobei die zur Sortierung verwendeten Kriterien nicht einsehbar und damit nur teilweise bekannt sind. Bei Flickr können die Suchresultate beispielsweise dadurch verbessert werden, dass bestimmte Autorinnen und Autoren als „untrustworthy" eingestuft und ihre Bilder bei Suchergebnissen herausgefiltert werden [AV09]. Flickr analysiert sogar die Bilder selbst und versucht über technische Kriterien wie Kontrast, Farbsättigung und Bildschärfe eine interne Qualitätsbewertung zur Rangierung vorzunehmen [PS09]. Auch bei Web-2.0-Bildungsplattformen für user-generated educational Microcontent sind spezifische und maschinell auswertbare Kriterien wünschenswert, um den Sortieralgorithmus und damit die Nutzenden bei der Suche nach qualitativ guten und relevanten Inhalten zu unterstützen.

Über Qualitätsaspekte von Lerninhalten gibt es diverse Kriterienkataloge und Analyseraster. Ein didaktisches Kategoriensystem zur Analyse des kognitiven Potenzials von Aufgaben wird beispielsweise von Maier et al. [Ma10] beschrieben. Ein mögliches Beurteilungsraster für die didaktische Qualität von Multimediaprodukten wurde von Mikuszeit & Szudra [MS09] vorgestellt. Darin wurden Qualitätskriterien, wie beispielsweise inhaltliche Korrektheit, klar erkennbares Lernziel oder Verständlichkeit der Aufgabenstellung, aufgestellt. Diese eignen sich jedoch nur eingeschränkt für die automatisierte Auswertung durch den Computer. Die didaktische Qualität eines Lernbausteins kann maschinell nicht abschließend beurteilt werden und erfordert eine manuelle Sichtung durch eine erfahrene Lehrperson. Vergleichbar mit der Analyse der Farbsättigung bei Bildern, die keine Aussage über die inhaltliche Qualität liefern kann, könnte bei einer interaktiven Zuordnungsübung aber zum Beispiel ermittelt werden, wie oft die Übung von Lernenden erfolgreich gelöst bzw. vorzeitig abgebrochen wurde. Über solche messbaren statistischen Größen lassen sich gewisse positive bzw. negative Tendenzen zur Qualität eines Inhalts ableiten. Eine GeoGebra-Simulation lässt sich zwar nur bedingt mit einem Kreuzworträtsel oder einer multimedialen Zuordnungsübung vergleichen, jedoch lassen sich ähnliche Aussagen zum Beispiel über deren Autorinnen und Autoren und das Konsumverhalten der Nutzenden machen. In Abbildung 3 sind typische Interaktionen zwischen Inhalt, Autor/in und Nutzenden schematisch dargestellt.

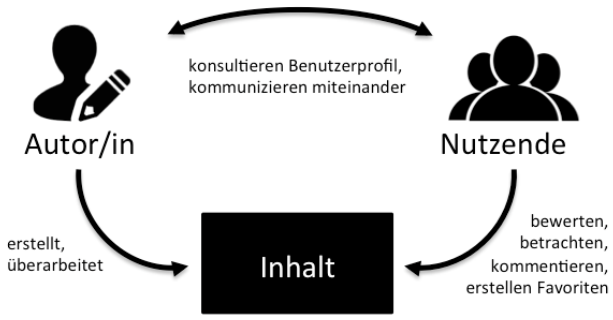

Abbildung 3: Interaktionen zwischen Autorin bzw. Autor, Inhalt und Nutzenden

Die im folgenden Kapitel beschriebenen Kriterien wurden jeweils mit Hilfe der Datenbasis von LearningApps.org auf einen statistischen Zusammenhang mit der manuellen Bewertung durch Nutzende untersucht. Die Untersuchungen basieren somit auf der Annahme, dass ein tatsächlicher Zusammenhang zwischen der Bewertung von Nutzenden und der didaktischen Qualität eines Inhalts, sowohl in positiver als auch negativer Richtung, besteht. Die Annahme wird unter anderem dadurch gestützt, dass die Bewertungen der Inhalte stark mit der Anzahl erhaltener Favoriten korreliert – und dies obwohl die Motivation zum Bewerten (Feedback für Autor/in oder indirekte Empfehlung für andere Nutzende) und dem Merken eines Inhalts als Favorit (für den späteren Eigengebrauch, typischerweise durch Lehrpersonen) sehr verschieden ist. Inhalte, die als Favorit vermerkt werden, entsprechen den Bedürfnissen der jeweiligen Lehrpersonen. Es kann davon ausgegangen werden, dass damit auch die didaktische Qualität dieser Inhalte adäquat ist.

Für Inhalte auf LearningApps.org wurde zum Zeitpunkt der Untersuchung ein Bewertungssystem mit 1 bis 5 Sternen eingesetzt. Die Auswertung von insgesamt über 14'000 von den Nutzenden abgegebenen Bewertungen zeigt eine Verteilung, die besonders die Extreme betont: 71% entfallen auf eine Bewertung mit fünf Sternen und 14% auf eine Bewertung mit nur einem Stern. Ebenso zeigte die Analyse, dass die durchschnittliche Bewertung aller Inhalte einer Autorin bzw. eines Autors tendenziell entweder positiv (4-5 Sterne) oder negativ (1-2 Sterne) ausfällt. Es kann somit vereinfacht auch von gut bzw. schlecht bewerteten Autorinnen und Autoren gesprochen werden. Von den rund 55'000 öffentlichen und damit bewertbaren Inhalten erhielten jedoch nur 10% eine oder mehrere Bewertungen.

Ziel der automatisierten Bewertung ist es, das Bewertungsverhalten der Nutzenden für bislang unbewertete Inhalte möglichst gut nachzuahmen. Ein Algorithmus berechnet anhand von Kriterien eine Kennzahl pro Inhalt, mit deren Hilfe eine Rangierung vorgenommen werden kann. Um die Gewichte der einzelnen Kriterien bei der Berechnung zu bestimmen, können die bereits bewerteten Inhalte als Trainings- und Validierungsdatensätze verwendet werden.

4 Kriterien und Evaluation

Im Rahmen dieser Arbeit wurde ein Kriterienkatalog für die automatisierte Bewertung von Lernbausteinen anhand statistischer Daten entwickelt. Die Aussagefähigkeit einer solchen Bewertung ist eng mit der vorhandenen Datenmenge verknüpft und wird somit erst bei Plattformen mit vielen tausenden Inhalten und Nutzenden möglich. Die Wahl der Kriterien erfolgte mit dem Anspruch, dass diese bei möglichst vielen Plattformen für user-generated educational Microcontent eingesetzt werden können. Im Folgenden werden die acht Kriterien beschrieben, die sich im Rahmen der Untersuchungen mit Daten von LearningApps.org als geeignet erwiesen haben.

Kriterium 1: Erhaltene Favoriten pro Aufruf

Bei vielen Web-2.0-Diensten besteht die Möglichkeit, einen Inhalt für den späteren Gebrauch als Favorit zu markieren. Das Setzen eines Favoriten kann als eine manuelle Sichtung und bewusste Selektion verstanden werden. Typischerweise werden Favoriten bei Bildungsplattformen durch Lehrpersonen während der Suche nach Materialien für den eigenen Unterricht gesetzt. Auf LearningApps.org wurde untersucht, ob ein Zusammenhang zwischen erhaltenen Favoriten und der durchschnittlichen Sterne-Bewertung eines Inhalts durch andere Nutzende besteht. Bei Abbildung 4 wurde die normierte Größe Anzahl Favoriten pro Aufruf betrachtet. Die Auswertung zeigt, dass gut bewertete Inhalte (4-5 Sterne) mit einem Korrelationskoeffizient von $r = 0{,}92$ auch öfters als Favorit gemerkt werden. Rund 70% aller mit 5 Sternen bewerteten Inhalte wurden mindestens einmal von anderen Nutzenden als Favorit markiert. Es wurden ca. 56'000 verschiedene Inhalte rund 140'000 mal als Favorit gespeichert. Insgesamt werden rund zehnmal mehr Inhalte als Favorit markiert als Bewertungen abgegeben. Die durchschnittliche Anzahl Favoriten pro Aufruf lässt sich als Bewertungskriterium für einen Inhalt nutzen.

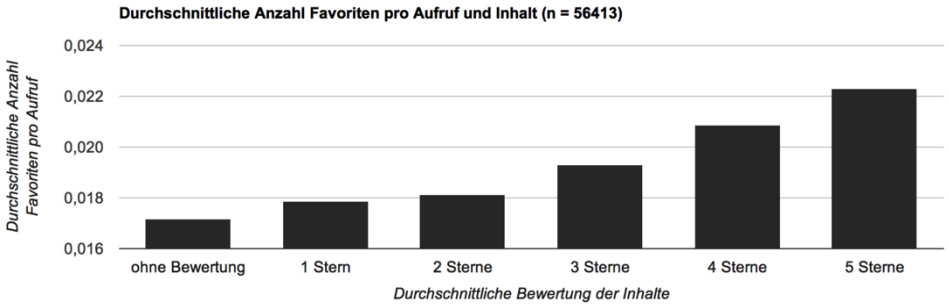

Abbildung 4: Durchschnittliche Anzahl Favoriten pro Aufruf und Inhalt

Kriterium 2: Anzahl Nutzungen eines Inhalts in Gruppen oder Klassen

Bei allen drei evaluierten Bildungsplattformen können Inhalte in Kollektionen zusammengestellt werden. Bei LearningApps.org haben Lehrpersonen die Möglichkeit, eine Kollektion pro Schulklasse zu erstellen, um Inhalte ihren Schülerinnen und Schülern zur Verfügung zu stellen. Insgesamt wurden rund 30'000 verschiedene Inhalte in 13'000 Schulklassen bereitgestellt. Es wurde analysiert, ob die Inhalte, die in vielen Kollektionen unterschiedlicher Lehrpersonen eingesetzt wurden, auch besser bewertet wurden und mehr Favoriten pro Aufruf erhalten haben. Die in Abbildung 5 dargestellten Resultate zeigen keinen signifikanten Zusammenhang zwischen der Anzahl Klassen und den Bewertungen, jedoch einen starken Zusammenhang mit den Favoriten auf (r = 0,98). Die Anzahl Kollektionen, in denen ein Inhalt verwendet wird, kann somit als Bewertungskriterium herbeigezogen werden.

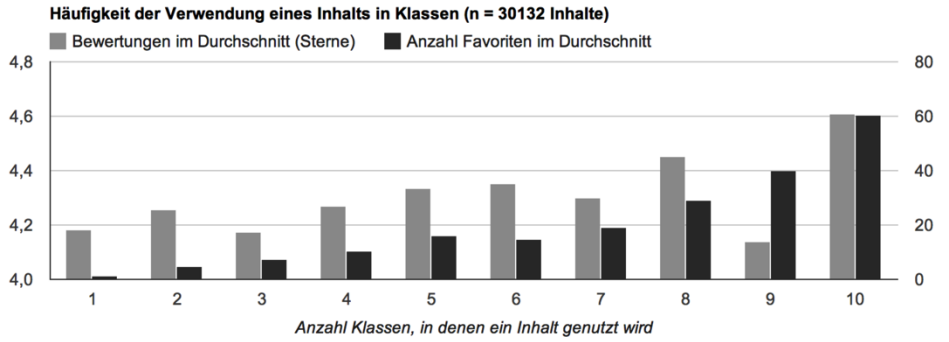

Abbildung 5: Anzahl Klassen und Inhalte

Kriterium 3: Anzahl Angaben im Benutzerprofil

Wenn Autorinnen und Autoren Inhalte auf einer Plattform erstellen und veröffentlichen wollen, müssen sie sich in der Regel registrieren und einen Benutzeraccount anlegen. Dabei werden im Benutzerprofil meistens Angaben zur Person erfasst. Häufig lassen sich neben den geforderten Pflichtfeldern wie Benutzername und Passwort weitere freiwillige Angaben, wie E-Mail-Adresse, Wohnort, Schule oder eigene Homepage, hinterlegen. Anhand der Daten von LearningApps.org wurde geprüft, ob ein Zusammenhang zwischen den freiwilligen Angaben im Benutzerprofil und der durchschnittlichen Bewertung bzw. Anzahl Favoriten, welche die Autorinnen und Autoren erhalten haben, be-

steht. Es wurden rund 100'000 Benutzerprofile untersucht. Für die Sterne-Bewertung zeigte sich kein signifikanter Zusammenhang. Unabhängig von den Angaben im Benutzerprofil betrug die durchschnittliche Bewertung aller Inhalte 3,9 ±0.2 Sterne. In Abbildung 6 ist jedoch ersichtlich, dass ein Zusammenhang mit der Anzahl erhaltener Favoriten besteht. Inhalte von Autorinnen und Autoren, die zusätzlich zum Namen noch weitere Angaben zur E-Mail, Website und Schule eingetragen haben, werden deutlich häufiger als Favorit gespeichert (r = 0,95). Das Eintragen bzw. Sichtbarstellen von zusätzlichen Angaben im Benutzerprofil kann somit als Bewertungskriterium für die Inhalte der Autorin bzw. des Autors genutzt werden.

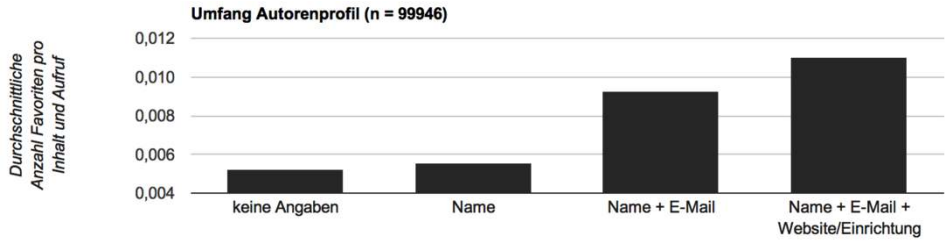

Abbildung 6: Vergleich Angaben im Benutzerprofil bei LearningApps.org

Kriterium 4: Kommunikation mit anderen Nutzenden

Nutzerinnen und Nutzer haben bei vielen Web-2.0-Diensten die Möglichkeit, einer Autorin oder einem Autor private Nachrichten zu senden. In Abbildung 7 wurde die durchschnittliche Anzahl geschriebener und erhaltener Nachrichten von insgesamt rund 110'000 Autorinnen und Autoren von LearningApps.org untersucht. Es wurde geprüft, ob ein Zusammenhang mit der durchschnittlichen Bewertung ihrer Inhalte und der Anzahl Nachrichten besteht. Es zeigt sich, dass Autorinnen und Autoren mit durchschnittlich gut bewerteten Inhalten (4-5 Sterne) überdurchschnittlich viel kommunizieren. Die absolute Anzahl gesendeter und empfangener Nachrichten der Autorin bzw. des Autors kann somit als Bewertungskriterium für dessen Inhalte verwendet werden.

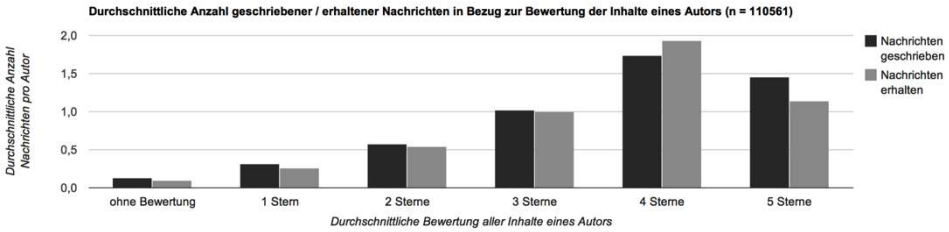

Abbildung 7: Durchschnittliche Anzahl Nachrichten pro Autor und Bewertung

Kriterium 5: Anzahl bereitgestellter Inhalte

Autorinnen und Autoren erstellen in der Regel mehrere Inhalte und stellen diese über die Plattform anderen Nutzenden zur Verfügung. Auf LearningApps.org wurde für rund 2'900 Autorinnen und Autoren untersucht, ob ein Zusammenhang zwischen der Anzahl der von ihnen bereitgestellten Inhalte und der durchschnittlich erhaltenen Bewertung

durch andere Nutzende besteht. Mit einem Korrelationskoeffizient r = 0,86 zeigt sich, dass Autorinnen und Autoren mit einer hohen Durchschnittsbewertung tendenziell mehr Inhalte erstellten als jene mit niedrigen Sterne-Bewertungen (siehe Abbildung 8). Die Gesamtzahl erstellter Inhalte einer Autorin bzw. eines Autors kann infolgedessen als Bewertungskriterium eines Inhalts genutzt werden.

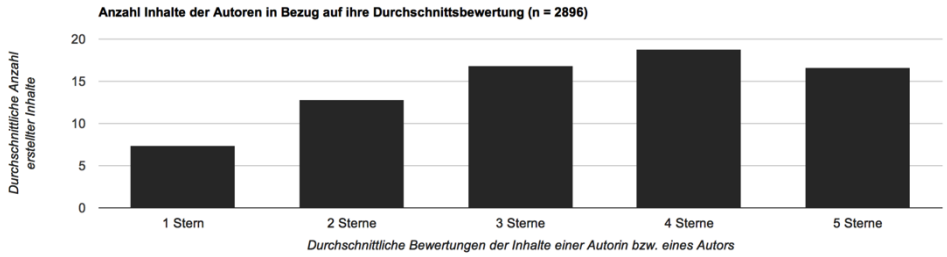

Abbildung 8: Durchschnittliche Anzahl erstellter Inhalte pro Autor und Bewertung

Kriterium 6: Meta-Informationen zu einem Inhalt

Autorinnen und Autoren können zu ihren erstellten Inhalten in der Regel zusätzliche Meta-Daten wie beispielsweise Tags, Informationen zur geeigneten Schulstufe oder zum Unterrichtsfach anfügen. Bei LearningApps.org wurde analysiert, ob gut bewertete Inhalte mehr Meta-Informationen aufweisen als schlechter bewertete Inhalte. Dazu wurde bei rund 57'000 öffentlichen Inhalten untersucht, ob sie Angaben zur Schulstufe sowie Tags zur Beschreibung des Inhalts enthalten. Abbildung 9 zeigt, dass Inhalte mit diesen zusätzlichen Meta-Daten tendenziell besser bewertet wurden (Korrelationskoeffizient r = 0,89). Die Angabe von Meta-Informationen kann somit für die Berechnung der Bewertungskennzahl eines Inhalts verwendet werden.

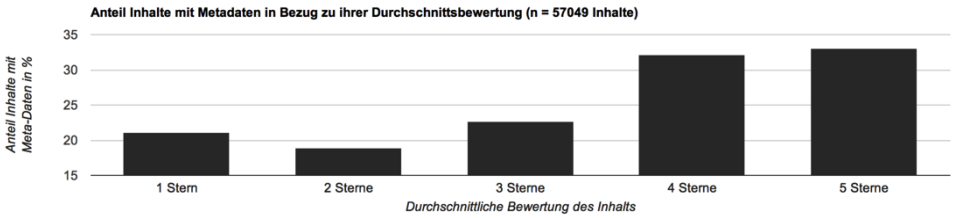

Abbildung 9: Metadaten bei Inhalten in Bezug zu ihrer Durchschnittsbewertung

Kriterium 7: Gründen von Gruppen oder Klassen

Einige Dienste bieten Autorinnen und Autoren die Möglichkeit, Gruppen oder Klassen zu erstellen. Eine Gruppe oder Klasse kann mehrere Mitglieder besitzen, die thematisch oder organisatorisch zusammen gehören. Bei LearningApps.org wurde untersucht, ob die Inhalte von Autorinnen und Autoren tendenziell besser bewertet werden, wenn sie eine oder mehrere Klassen erstellt haben. Betrachtet man alle Autorinnen und Autoren auf LearningApps.org und setzt die durchschnittliche Bewertung ihrer Inhalte mit der Grün-

dung von Klassen ins Verhältnis, ergibt sich eine Verteilung gemäß Abbildung 10. Rund 20% der Autorinnen und Autoren erstellten mindestens eine Klasse. Es zeigt sich, dass besonders Autorinnen und Autoren mit gut bewerteten Inhalten auch mehr Klassen erstellen. Die Anzahl gegründeter Klassen einer Autorin bzw. eines Autors korreliert mit einem Korrelationskoeffizient von r = 0,75 mit der durchschnittlichen Bewertung durch andere Nutzende und kann daher als Bewertungskriterium für Inhalte der Autorin bzw. des Autors genutzt werden.

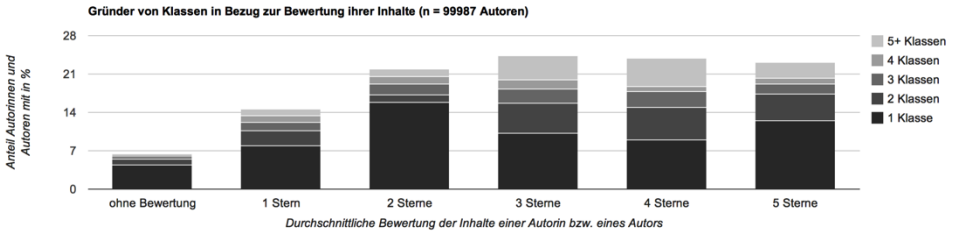

Abbildung 10: Gründung von Klassen in Bezug zur durchschnittlichen Bewertung

Kriterium 8: Anzahl Aufrufe des Benutzerprofils

Die Benutzerprofile bieten anderen Nutzenden Hinweise zu einer Autorin bzw. zu einem Autor. Anhand eines Benutzerprofils lässt sich manchmal mehr über das Tätigkeitsgebiet oder die Interessen der Autoren erfahren, was wiederum Hinweise auf die Qualität der erstellten Inhalte geben könnte. Auf LearningApps.org wurde untersucht, ob Autorinnen und Autoren, bei denen das Benutzerprofil häufig aufgerufen wurde, durchschnittlich bessere Bewertungen für ihre Inhalte durch andere Nutzende erhalten haben. Abbildung 11 zeigt, dass die Inhalte bei Autoren, deren Benutzerprofil weniger häufig aufgerufen wurde, tendenziell schlechter bewertet wurden. Mit einem Korrelationskoeffizient von r = 0,75 lässt sich ein Zusammenhang bestätigen. Die tägliche Anzahl Aufrufe des Benutzerprofils der Autorin bzw. des Autors kann folglich als Bewertungskriterium eines Inhalts verwendet werden.

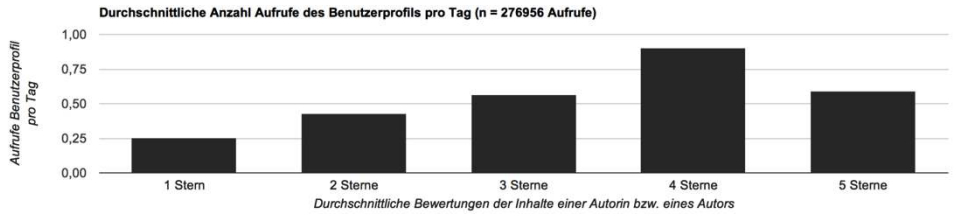

Abbildung 11: Aufrufe Benutzerprofil pro Tag in Bezug zu den durchschnittlichen Bewertungen

Kriterien ohne erkennbare Relevanz

Anhand der Daten von LearningApps.org wurde auch untersucht, ob ein Zusammenhang zwischen der Bewertung durch Nutzende und:

- der durchschnittlichen Anzahl Aufrufe pro Tag eines Inhalts
- der Anzahl Überarbeitungen eines Inhalts durch die Autorin bzw. den Autor
- dem Anteil öffentlich bereitgestellter Inhalte gegenüber privaten Inhalten einer Autorin bzw. eines Autors
- dem Anteil Aufrufe mit vollständiger Lösung (ob ein Lerninhalt bis zu Ende gelöst wurde) bzw. frühzeitigen Abbruch durch Lernende
- der durchschnittlichen Bearbeitungszeit der Lernenden

besteht. Für alle diese Kriterien ließ sich jedoch keine eindeutige Korrelation ermitteln. Zum Beispiel könnte analog zu Portalen wie YouTube erwartet werden, dass die Dauer der Auseinandersetzung mit dem Inhalt bis zur vollständigen Lösung der Aufgabe ein relevantes Kriterium sei. In Abbildung 12 zeigt sich jedoch, dass keine signifikante Korrelation zwischen der Bewertung der Inhalte mit Sternen und der durchschnittlichen Bearbeitungszeit durch Lernende festzustellen ist.

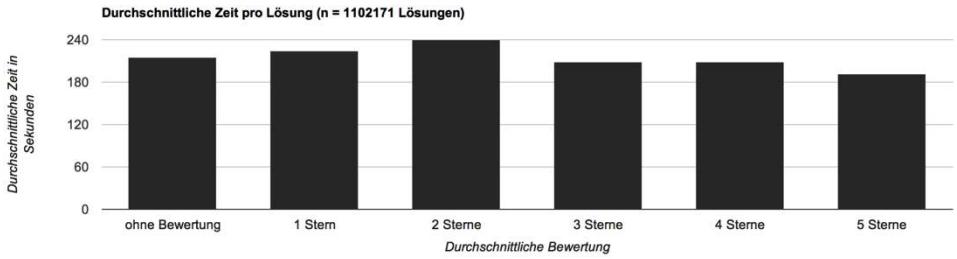

Abbildung 12: Durchschnittliche Bearbeitungszeit bis zur Lösung eines Inhalts

Ob diese Kriterien für die automatisierte Bewertung generell ungeeignet sind, oder ob die vorhandene Datenbasis unzureichend war, bleibt Gegenstand für nachfolgende Untersuchungen.

5 Anwendung der Kriterien am Beispiel LearningApps.org

Anhand der acht als relevant ermittelten Kriterien konnte eine einfache Bewertungsfunktion mit Hilfe linearer Regression als gewichtete Summe der Kriterienwerte aufgestellt werden. Die Funktion liefert pro Inhalt eine Bewertung zwischen 1 und 5, die möglichst gut mit der Sternebewertung der Nutzenden übereinstimmen soll. Die Gewichte wurden mit Hilfe einer Trainingskollektion von rund 1000 mehrfach bewerteten Inhalten (rund 200 Inhalte pro Sternekategorie) bestimmt. Die Bewertungsfunktion wurde anschließend anhand einer Testkollektion aus weiteren 1000 Inhalten mit ähnlich verteilten Benutzerbewertungen validiert. Die Bewertungsfunktion stimmt mit einem mittleren Fehler von 0,84 Sternen und einem Korrelationskoeffizienten von r = 0,67 mit der Sternebewertung

durch die Nutzenden überein. In Abbildung 13 wurde die Abweichung der algorithmischen Bewertung pro Sternekategorie von der Bewertung durch Nutzende dargestellt. Vereinfacht man die Bewertung und betrachtet 3 Sterne als neutral, alle Inhalte mit weniger als 3 Sternen als negativ und die Inhalte mit mehr als 3 Sternen als positiv, konnten 363 von 468 (78%) der positiven und 471 von 591 (79%) der negativen Inhalte der Testkollektion entsprechend richtig zugeordnet werden. Die Bewertungsfunktion wurde anschließend auf rund 50'000 öffentliche Inhalte angewendet, die bislang noch keine Bewertung durch Nutzende erhalten hatten. Die Verteilung der Bewertungen ist in Abbildung 14 dargestellt. Der Algorithmus bewertete den überwiegenden Teil mit einer neutralen 3-Sterne-Bewertung. Ein Drittel der Inhalte erhielt eine positive und rund 5% der Inhalte eine negative Bewertung. Im Vergleich zum Rangierungsverfahren ausschließlich auf Basis der Bewertungen durch Nutzende, bei dem auf Grund der geringen Anzahl Bewertungen über 90% der Inhalte unsortiert bleiben, ist dies eine Verbesserung. Die Bewertungsfunktion wurde auf LearningApps.org als neue Rangierungsvariante bei der Sichtung von Inhalten eingefügt und steht allen Nutzenden zur Verfügung.

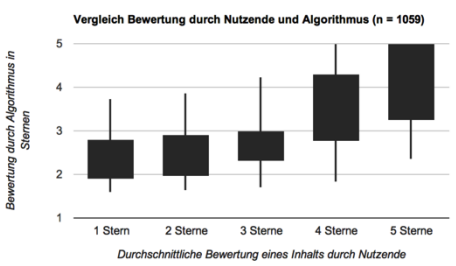

Abbildung 13: Abweichung Bewertungsfunktion

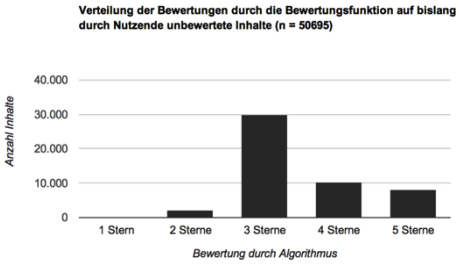

Abbildung 14: Verteilung unbewertete Inhalte

6 Fazit und Ausblick

Die kaum vorhandene Qualitätsprüfung und Bewertung von Open Educational Resources auf Web-2.0-Bildungsplattformen macht die Erschließung geeigneter Inhalte für Lehrpersonen schwierig. Gleichzeitig macht die Menge an Inhalten eine manuelle Prüfung durch Expertinnen und Experten unmöglich. Im Rahmen dieser Arbeit wurden mehrere Plattformen für freie und interaktive Lernbausteine untersucht und die Interaktionen zwischen Inhalt, Autor/in und Nutzenden analysiert. Mit Hilfe einer automatisierten Bewertung wurde versucht, das Bewertungsverhalten von Nutzenden auf bislang unbewertete Inhalte zu übertragen. Dazu wurden insgesamt 13 Bewertungskriterien anhand statistischer Daten aufgestellt. Die Kriterien wurden mit Hilfe der Plattform LearningApps.org evaluiert und acht Kriterien als geeignet identifiziert. Der entstandene Kriterienkatalog unterstützt die Neu- und Weiterentwicklung von Bildungsplattformen für user-generated educational Microcontent. Durch eine bessere Rangierung und Filterung der Inhalte können Lehrpersonen rascher und effizienter passende Inhalte finden, wodurch sich der Wert der Plattform und die Zufriedenheit der Nutzenden steigern lässt. Die automatisierte Bewertung anhand statistischer Daten kann jedoch nur als zusätzliches Hilfsmittel eingesetzt werden und erfordert weiterhin eine aktive Bewertung der Inhalte durch die Nutzenden selbst. Die Qualität der algorithmischen Bewertung ist

zudem nur schwer zu beurteilen. Um den tatsächlichen Zusammenhang der Bewertungen der Nutzenden bzw. eines Algorithmus mit der didaktischen Qualität eines Lernbausteins zu erheben, wären umfangreiche Vergleichsstudien mit erfahrenen Lehrpersonen notwendig. In einem nächsten Schritt ist deshalb eine Pilotstudie vorgesehen, in welcher Expertinnen und Experten in einem Themenfeld eine vorgegebene Kollektion von Inhalten rangieren sollen und diese Rangierung anschließend mit der automatischen Rangierung verglichen wird. Die Rangierung der Inhalte könnte auch durch Collaborative Filtering und Recommender Systems Algorithmen [Sc07], welche die Interessen und Vorlieben der Nutzenden erfassen und vergleichen, weiter verbessert werden.

Literaturverzeichnis

[AV09] Allan, M.; Verbeek, J. J.: Ranking User-annotated Images for Multiple Query Terms. In *BMVC 2009 - British Machine Vision Association*, 2009

[Ge07] Geser, G.: Open Educational Practices and Resources: OLCOS Roadmap 2012. Open eLearning Content Observatory Services (OLCOS). Salzburg Research, EduMedia Group, 2007.

[Hi13] Hielscher, M.: Autorentools für multimediale und interaktive Lernbausteine: Architektur und Einsatzszenarien von LearningApps.org. Hülsbusch, 2013.

[HHR13] Hielscher M.; Hartmann W.; Rothlauf F.: Entwicklung eines Autorenwerkzeuges für digitale, multimediale und interaktive Lernbausteine im Web 2.0. In *DeLFI 2013, E-Learning Fachtagung Informatik, Lecture Notes in Informatics (LNI) - Proceedings. Series of the Gesellschaft für Informatik (GI).* Volume P-218, S. 203-214, 2013

[Ma10] Maier, U.; Kleinknecht, M.; Metz, K.; Bohl, T.: Ein allgemeindidaktisches Kategoriensystem zur Analyse des kognitiven Potenzials von Aufgaben. Beiträge zur Lehrerbildung. 28(1), S. 84–96. 2010

[MS09] Mikuszeit, B.; Szudra, U.: Multimedia und ethische Bildung: E-Learning - Ethik - Blended-Learning. Peter Lang Publishing Group, 2009.

[PS09] Pedro, J. S.; Siersdorfer, S.: Ranking and classifying attractiveness of photos in folksonomies. In *WWW '09: Proceedings of the 18th international conference on World wide web*, S. 771-780. ACM, 2009.

[Si10] Siersdorfer, S.; Chelaru, S.; Nejdl, W.; Pedro, J. S.: How Useful are Your Comments? Analyzing and Predicting YouTube Comments and Comment Ratings. In *WWW '10, Proceedings of the 18th international conference on World wide web*, S. 891-900, ACM, 2010

[Sc07] Schafer, J. B.; Frankowski, D.; Herlocker, J.; Sen, S.: Collaborative Filtering Recommender Systems. Methods and Strategies of Web Personalization. In The Adaptive Web, S. 291-324, 2007