Automatische Reputationsmessung in der Wikipedia

Thomas Wöhner
Martin-Luther-Universität
Halle-Wittenberg
06099 Halle (Saale)
+49 345 55-23478
thomas.woehner@wiwi.uni-halle.de

Sebastian Köhler
Martin-Luther-Universität
Halle-Wittenberg
06099 Halle (Saale)
+49 345 55-23479
sebastian.koehler@wiwi.uni-halle.de

Ralf Peters
Martin-Luther-Universität
Halle-Wittenberg
06099 Halle (Saale)
+49 345 55-23471
ralf.peters@wiwi.uni-halle.de

ZUSAMMENFASSUNG

Die Wikipedia wird im Rahmen ihres offenen Zugangs von Autoren erstellt, die überwiegend anonym oder unter Pseudonym tätig und damit weitgehend unbekannt sind. Daraus resultiert eine Unsicherheit bezüglich der Güte der einzelnen Beiträge. Einen Lösungsansatz bieten automatische Reputationssysteme, die sich in den vergangenen Jahren als eigenständiges Forschungsgebiet etabliert haben. Durch diese Reputationssysteme wird die Reputation der Autoren anhand ihrer bisherigen Bearbeitungen automatisch berechnet. In der gegenwärtigen Forschung werden die zur Reputationsmessung vorgeschlagenen Metriken jedoch nur isoliert betrachtet und oft unzureichend bewertet, sodass sich deren Aussagekraft nur schwer abschätzen lässt. Im vorliegenden Beitrag werden insgesamt zehn Metriken vergleichend bewertet und durch Kombination anhand einer Diskriminanzanalyse zu einer effizienten Reputationsfunktion zusammengeführt. Die Metriken wurden der bestehenden Literatur entnommen und um eigene Vorschläge ergänzt. Die Analyse zeigt, dass die neu vorgeschlagene Metrik der *Effizienz der Bearbeitungen* besonders aussagekräftig ist.

Schlüsselwörter

Web 2.0, Wikipedia, Reputation, Reputationssystem, Persistenz, Qualität, Effizienz

1. EINLEITUNG

Unter dem Begriff Web 2.0 gewannen in den vergangenen Jahren Websites mit benutzergenerierten Inhalten (User Generated Content) zunehmend an Bedeutung [17]. Die besondere Relevanz des Web 2.0 zeigt sich beispielsweise an der Besucherstatistik der zehn weltweit meistbesuchten Websites, von denen fünf (Facebook, YouTube, Wikipedia, Blogger, Twitter) auf benutzergenerierten Inhalten basieren.[1] Ein weit verbreiteter Anwendungstyp innerhalb des Web 2.0 sind Wikis. Wikis sind Websites, deren Inhalt direkt im Browser durch die Internetbenutzer geändert werden kann. Die entsprechende Wiki-Syntax ist vergleichsweise einfach, sodass eine Bearbeitung ohne besondere technische Vorkenntnisse möglich ist. Wikis werden daher zur kollaborativen Erstellung von Websites eingesetzt [6].

Das weltweit größte und bekannteste Wiki ist die freie Online-Enzyklopädie Wikipedia, die im Januar 2010 mehr als *15* Millionen Artikel[2] enthielt und in mehr als *260* Sprachen[3] verfügbar ist. Entsprechend dem Wiki-Prinzip werden die Artikel ausschließlich durch die Internetbenutzer erstellt und unterliegen keiner Kontrolle durch Experten. Jede Änderung wird unmittelbar im World Wide Web veröffentlicht. Dieses offene Konzept führt auf der einen Seite dazu, dass sich viele Internetbenutzer an der Erstellung und Pflege der Wikipedia beteiligen. Durch die sogenannte *Weisheit der Vielen* [22] erfolgt eine zeitnahe Aktualisierung und Fehler werden zumeist schnell erkannt und korrigiert. Studien haben ergeben, dass die Wikipedia ein mit klassischen Enzyklopädien wie Brockhaus oder Britannica vergleichbares Qualitätsniveau erreicht [9, 13]. Auf der anderen Seite lassen sich durch die beschriebene Offenheit unerwünschte Änderungen wie Vandalismus, Spam und fehlerhafte Einträge aufgrund von Unwissenheit und Opportunismus nicht ausschließen [7, 25, 26]. Da die Bearbeitungen in der Wikipedia von weitgehend unbekannten Autoren vorgenommen werden, ist die Güte der einzelnen Beiträge nur schwer einzuschätzen.

Ein weit verbreiteter Ansatz, um Vertrauen zwischen unbekannten Teilnehmern zu schaffen, sind Reputationssysteme [20]. Reputationssysteme erfassen und bewerten das Verhalten der Teilnehmer in der Vergangenheit, um darauf aufbauend deren zukünftiges Verhalten abzuschätzen. Man unterscheidet dabei zwischen expliziten (benutzergetriebenen) und impliziten (automatischen) Reputationssystemen. Bei expliziten Reputationssystemen wird die Bewertung von den Nutzern des jeweiligen Systems abgegeben. Ein bekanntes Beispiel hierfür ist das Reputationssystem des Online-Auktionshauses eBay. Eine explizite Reputationsbewertung ist in der Wikipedia jedoch nicht praktikabel, da die MediaWiki-Software nicht darstellt, von welchem Autor ein Artikel bzw. einzelnen Textabschnitte erstellt wurden. Zur Ermittlung der Autorenschaft als Voraussetzung für eine benutzergetriebene Bewertung müsste daher zunächst die MediaWiki-Software erweitert oder die Versionsgeschichte manuell ausgewertet werden.

[1] http://www.alexa.com/topsites
[2] http://stats.wikimedia.org/DE/TablesWikipediaZZ.htm
[3] http://de.wikipedia.org/wiki/Wikipedia:Sprachen

10th International Conference on Wirtschaftsinformatik,
16th - 18th February 2011, Zurich, Switzerland

Im Kontext der Wikipedia haben sich in den vergangenen Jahren insbesondere implizite Reputationssysteme als eigenständiges Forschungsgebiet etabliert [1, 11]. Bei impliziten Reputationssystemen wird die Reputation automatisch berechnet. Im Falle der Wikipedia wird dabei durch Auswertung der Versionsgeschichte auf das Bearbeitungsverhalten der Autoren geschlossen und darauf aufbauend die Reputation errechnet. Die Schwierigkeit besteht dabei in der Auswahl geeigneter Metriken, die das typische Bearbeitungsverhalten von guten bzw. schlechten Autoren effektiv erfassen.

Reputationssysteme können in der Wikipedia unterschiedliche Funktionen erfüllen. Anhand des Reputationswertes lässt sich beispielsweise die Güte der Bearbeitungen eines Autors abschätzen [2] oder ein Qualitätsscore für gesamte Artikel berechnen. Auch kann das Reputationssystem genutzt werden, um Bearbeitungsrechte in der Wikipedia adäquat zu beschränken [1]. Des Weiteren kann das Reputationssystem die Autoren motivieren, sich mit qualitativ hochwertigen Bearbeitungen intensiv an der Wikipedia zu beteiligen [11].

Die Forschung zur automatischen Reputationsmessung in der Wikipedia steht bisher noch am Anfang. In den gegenwärtigen Publikationen werden bereits einzelne Metriken zur Reputationsmessung vorgeschlagen. Teilweise fehlt es aber an einer Bewertung der vorgeschlagenen Metriken. Darüber hinaus werden die einzelnen Metriken bisher nur isoliert untersucht.

Der vorliegende Beitrag adressiert diese Forschungslücke und untersucht zum einen zehn potentielle Metriken zur automatischen Reputationsmessung daraufhin, welche Einzelmetrik die größte Aussagekraft zur Reputationsmessung besitzt. Zum anderen werden die Metriken kombiniert und zu einer Reputationsfunktion zusammengeführt. Dahinter steht die Idee, dass durch eine solche Kombination verschiedene Facetten des Bearbeitungsverhaltens erfasst werden können. Dabei werden sowohl neu entwickelte als auch aus der Literatur bekannte Metriken berücksichtigt.

Zur Evaluation werden die Metriken anhand einer Diskriminanzanalyse zur Klassifikation zwischen schlechten (gesperrte Benutzer) und guten (nicht gesperrte Benutzer) Autoren in der Wikipedia herangezogen. Die Diskriminanzanalyse liefert die Reputationsfunktion mit maximaler Klassifikationsgüte, in die die Metriken linear mit einem Gewichtungsfaktor eingehen. Die ermittelten Gewichtungskoeffizienten spiegeln damit auch die Aussagekraft der einzelnen Metriken bei der Reputationsmessung wider.

Der Beitrag ist wie folgt aufgebaut. In Kapitel 2 wird der derzeitige Stand der Forschung zur automatischen Reputationsmessung in der Wikipedia vorgestellt. Die Beschreibung der untersuchten Metriken erfolgt in Kapitel 3. In Kapitel 4 wird die Evaluationsmethodik detailliert dargestellt. Nachfolgend werden im Kapitel 5 die Metriken anhand einer Diskriminanzanalyse evaluiert und kombiniert. Abschließend erfolgt in Kapitel 6 die Schlussbetrachtung.

2. STAND DER FORSCHUNG

Hinsichtlich der Qualitätsproblematik in der Wikipedia sind in den vergangenen Jahren zahlreiche Publikationen entstanden. In den Arbeiten von Potthast et al. [18], Smets et al. [21], Priedhorsky [19] und West et al. [27] wird Vandalismus in der Wikipedia untersucht. Bei der Vandalismusdetektion wird beispielsweise anhand der Bearbeitungskommentare oder durch den Vergleich von Hashwerten analysiert, inwieweit Bearbeitungen wieder rückgängig gemacht werden. Viégas et al. [25, 26] zeigen, dass Vandalismus in der Wikipedia in der Regel innerhalb von drei Minuten korrigiert wird.

Zahlreiche andere Publikationen untersuchen Metriken zur automatischen Qualitätsbewertung der Artikel in der Wikipedia [5, 8, 14, 15, 16, 24, 29, 30]. Als besonders effiziente Qualitätsindikatoren gelten die Länge eines Artikels [5] und Lebenszyklus-basierte Messgrößen wie der Umfang der durchschnittlichen persistenten (effektiven) Änderungen [16, 29].

Während die Forschung zur automatischen Qualitätsbewertung der Artikel bereits weit fortgeschritten ist und effiziente Metriken erarbeitet wurden, steht die automatische Reputationsmessung in der Wikipedia noch am Anfang. Anthony et al. untersuchen registrierte und anonyme Nutzer anhand der Bearbeitungshäufigkeit (Anzahl der Bearbeitungen) sowie anhand des Umfangs und der Effizienz (prozentualer Anteil der Bearbeitungen, die in der neusten Version des Artikels enthalten sind) der Bearbeitungen [3]. Stein und Hess analysieren als Reputationsmetrik die Mitarbeit des Autors bei der Erstellung von qualitativ hochwertigen Artikeln [23]. In den Publikationen von Adler et al. [1, 2] und Javanmardi et al. [11] werden zwei einander ähnliche Ansätze zur Reputationsmessung vorgeschlagen, die sich entsprechend der Ausführungen von Javanmardi et al. [11] hinsichtlich der Komplexität der Berechnung unterscheiden. Beide stellen jeweils eine Reputationsmetrik vor, die auf Grundlage der Persistenz von Bearbeitungen berechnet wird. Dabei bemisst sich die Persistenz anhand der Zeitspanne, in der eine Änderung in der Wikipedia Bestand hat.

3. METRIKEN ZUR REPUTATIONSMESSUNG

Im folgenden Abschnitt wird zunächst ein formales Modell der Wikipedia entworfen. Darauf aufbauend werden potentielle Metriken zur Reputationsmessung beschrieben und deren Berechnung anhand des formalen Wikipediamodells erläutert.

3.1 Modell der Wikipedia

Die Wikipedia besteht aus einer Menge von Artikeln $i = 0 \ldots n$. Bei der Bearbeitung eines Artikels i durch einen Benutzer entsteht eine jeweils neue Artikelversion $v_{i,j}$ mit $j = 0 \ldots m$ (Anzahl der Versionen eines Artikels), welche in der Versionsgeschichte abgespeichert wird. Die erste Version $v_{i,0}$ eines Artikels i wird in unserem Modell als leeres Dokument definiert. Diese Annahme weicht von der tatsächlichen Datenbank der Wikipedia ab, in der die erste Version bereits den Titel des Artikels und die erste Bearbeitung enthält. Diese erste Version entspricht der Version $v_{i,1}$ in unserem Modell. Die Definition einer vorgehenden Version $v_{i,0}$ als leere Version ist erforderlich, da so die Bearbeitung des ersten Autors im Modell formal korrekt erfasst wird. Der Autor $editor(v_{i,j})$ der Version $v_{i,j}$ bezeichnet den Benutzer, durch dessen Bearbeitung die neue Version entstanden ist.

Die im Folgenden erläuterten Metriken basieren überwiegend auf dem Unterschied zwischen verschiedenen Versionen eines Artikels. Die Ermittlung von Textdifferenzen kann prinzipiell mit unterschiedlicher Granularität erfolgen. Man kann zwischen Textdifferenzberechnungen auf Zeilen-, Wort- oder Zeichenebene unterscheiden. Je feingranularer die Textberechnung erfolgt, desto rechenaufwändiger ist die Kalkulation. In Analogie zu anderen Arbeiten der Wikipediaforschung [1, 11] werden die Textunterschiede in der vorliegenden Analyse auf Wortebene berechnet.

Dabei wird ein Wort als eine Zeichenfolge zwischen zwei Leerzeichen definiert. Die Änderung eines einzelnen Buchstabens eines Wortes wird somit als Löschen und Hinzufügen eines Wortes interpretiert. Die Berechnung der Textvergleiche erfolgt mit dem weit verbreiteten Algorithmus von Hunt und McIlroy [10], der beispielsweise im Linux-Programm diff implementiert ist. Der Algorithmus basiert auf der Zerlegung des Textes in einzelne Token. Hunt und McIlroy definieren in ihrer Publikation die einzelnen Zeilen eines Textes (Lines of Code) als Token und ermitteln so den Textunterschied auf Zeilenebene. In der vorliegenden Analyse werden zur Verbesserung der Genauigkeit die einzelnen Wörter des Textes als Token verwendet.

Der Differenztext *del(i,j,z)* aus der Version $v_{i,j}$ und einer früheren Version $v_{i,z}$ mit $z < j$ enthält alle zwischenzeitlich gelöschten Textabschnitte. Analog zu *del(i,j,z)* beinhaltet der Differenztext *add(i,j,z)* alle neu hinzugefügten Token. Die Differenz *diff(i,j,z)* gibt den Gesamtunterschied wieder und umfasst sowohl *del(i,j,z)* als auch *add(i,j,z)*. Die Differenz *diff(i,j,j-1)* bezeichnet den Unterschied einer Version $v_{i,j}$ zur direkten Vorgängerversion $v_{i,j-1}$ und quantifiziert damit die Bearbeitung des Autors *editor*($v_{i,j}$). Der gemeinsame Text zweier beliebiger Texte t_1 und t_2 wird mit *equal*(t_1,t_2) bestimmt. Die Wortanzahl eines beliebigen Textes t wird mit $|t|$ bezeichnet.

3.2 Metriken zur Reputationsmessung

Die in diesem Beitrag untersuchten Metriken zur Reputationsmessung lassen sich in die Kategorien Bearbeitungshäufigkeit, Gesamtbearbeitungsumfang, Persistenz, Umfang der Bearbeitungen, Beteiligung an Diskussionen und Beteiligung an qualitativ hochwertigen Artikeln untergliedern. Die Metriken sind entweder aus der bestehenden Literatur entnommen oder werden auf Grundlage der Ziele der Wikipedia neu vorgeschlagen. Bei der Bestimmung der Messgrößen werden mit Ausnahme der in Kapitel 3.2.5 vorgestellten Metrik nur Bearbeitungen im Artikelnamensraum berücksichtigt. Dieser Namensraum enthält die enzyklopädischen Artikel der Wikipedia. Bearbeitungen anderer Namensräume, wie Bearbeitungen an Benutzerseiten oder Textvorlagen, tragen nur indirekt zum Fortschritt der Wikipedia bei und werden deshalb von der Untersuchung ausgeschlossen. Tabelle 1 listet die im Beitrag untersuchten Metriken auf.

3.2.1 Bearbeitungshäufigkeit

Die Bearbeitungshäufigkeit ist ein vergleichsweise einfaches Kriterium, das beispielsweise von Anthony et al. [3] diskutiert wird. Es quantifiziert den Gesamtbeitrag eines Autors zur Wikipedia und bestimmt damit seine Erfahrung. Anthony et al. [3] definieren hierzu als Messgröße die Anzahl

$$(1) \quad N^{e}_{a}$$

der Bearbeitungen eines Autors *a*. Im vorliegenden Beitrag wird als weitere potentielle Messgröße die Anzahl

$$(2) \quad N^{p}_{a}$$

der Artikel, die durch den Autor *a* bearbeitet wurden, vorgeschlagen.

3.2.2 Gesamtbearbeitungsumfang

Ähnlich der Bearbeitungshäufigkeit ist der Gesamtumfang der Bearbeitungen eine neu vorgeschlagene, potentielle Messgröße zur Bestimmung der Autorenreputation, die die Erfahrung des Autors in der Wikipedia erfasst. Dieser Umfang lässt sich anhand der Differenzen aller Versionen eines Autors zur jeweiligen Vorgängerversion bestimmen. Die Metrik

$$(3) \quad N^{w}_{a} = \sum_i \sum_j diff(i,j,j\text{-}1) \text{ mit } editor(v_{i,j}) = a$$

gibt die Anzahl der insgesamt von einem Autor *a* geänderten Wörter wieder.

Tabelle 1: Übersicht der untersuchten Metriken

Symbol	Bezeichnung	Quelle
$\mathbf{N^{e}_{a}}$	Anzahl der Bearbeitungen	[3]
$\mathbf{N^{p}_{a}}$	Anzahl der bearbeiteten Artikel	eigene Metrik
$\mathbf{N^{w}_{a}}$	Anzahl der insgesamt geänderten Wörter	eigene Metrik
$\mathbf{N^{pw}_{a}}$	Anzahl der persistent geänderten Wörter	eigene Metrik
$\mathbf{E_{a}}$	Effizienz	eigene Metrik
$\mathbf{Avg^{w}_{a}}$	durchschnittlicher Umfang der Bearbeitungen	[3]
$\mathbf{Avg^{pw}_{a}}$	durchschnittlicher Umfang der persistenten Änderungen	eigene Metrik
$\mathbf{max^{pw}_{a}}$	Umfang der größten persistenten Änderung	eigene Metrik
$\mathbf{N^{diss}_{a}}$	Anzahl der Bearbeitungen auf Diskussionsseiten	[12]
$\mathbf{N^{qh}_{a}}$	Beteiligung an qualitativ hochwertigen Artikeln	[23]

3.2.3 Persistenz

Die in der Literatur am häufigsten vorgeschlagenen Metriken zur Berechnung der Autorenreputation basieren auf der Persistenz der Bearbeitungen eines Autors [1, 3, 11]. Bei diesen Metriken wird bei einer hohen Verweildauer auf qualitativ hochwertige Änderungen geschlossen, da minderwertige Bearbeitungen frühzeitig von anderen Benutzern korrigiert werden [25, 26]. Bei der Berechnung der Persistenz ist sowohl das Einfügen als auch das Löschen von Text zu berücksichtigen, da auch das Löschen zur Verbesserung eines Artikels beitragen kann. Beispiele hierfür sind die Korrektur von Vandalismus und Spam-Einträgen.

In der Literatur werden unterschiedliche Berechnungsverfahren für die Persistenz der Bearbeitungen diskutiert, die jedoch verschiedene Nachteile aufweisen. Die Berechnungsverfahren von Adler et al. [1] und Javanmardi et al. [11] sind sehr rechenintensiv, da bei einer gegebenen Bearbeitung ein Textvergleich mit vielen Versionen durchzuführen ist. In der Publikation von Anthony et al. [3] wird die Persistenz daran bemessen, inwieweit die Bearbeitung in der jeweils neusten Version des Artikels enthalten ist. Diese Vorgehensweise weist den Nachteil auf, dass Löschungen nicht erfasst und somit nicht als persistente Änderungen erkannt werden.

In diesem Beitrag wird daher nicht auf eine Metrik aus der Literatur zurückgegriffen, sondern ein neuer Ansatz zur Bestimmung der Persistenz vorgeschlagen, der die Nachteile der bestehenden Verfahren vermeidet. Der Ansatz orientiert sich an den Überlegungen von Wöhner und Peters [29]. Wöhner und Peters analysieren in ihrer Publikation den Lebenszyklus von Artikeln und bezeichnen eine Änderung als persistent, falls die getätigte Bearbeitung bis zum Monatsende im Artikel verbleibt. Problematisch

hierbei ist jedoch, dass Änderungen am Ende eines Monats nur eine kurze Zeitspanne überdauern müssen und daher mit hoher Wahrscheinlichkeit als persistent klassifiziert werden. Aus diesem Grund wird im vorliegenden Beitrag als Persistenzkriterium eine einheitliche Mindestverweildauer verwendet, in der die Änderung nicht revidiert wird.

Bei der Festlegung der Mindestverweildauer sind zwei Aspekte zu berücksichtigen. Zum einen ist die Zeitspanne hinreichend groß zu wählen, damit unerwünschte Änderungen von der Wikipedia-Community innerhalb der Mindestverweildauer erkannt und revidiert werden. Das Erkennen und die Korrektur solcher unerwünschten Änderungen erfolgt in der Wikipedia in der Regel bereits innerhalb von drei Minuten [25, 26]. Anderseits darf die Zeitspanne nicht zu groß sein, da ansonsten wünschenswerte Änderungen im Rahmen der normalen Dynamik eines Artikels fälschlicherweise als nicht-persistent klassifiziert werden könnten.

Ausgehend von dieser Überlegung wurden Mindestverweildauern von einem Tag, zwei Tagen, zwei Wochen und zwei Monaten getestet, wobei die auf Basis der Persistenz berechneten Metriken nur geringfügig variierten. Die konkrete Ausgestaltung der Mindestverweildauer ist damit unkritisch für die ermittelten Ergebnisse. Die im Beitrag dargestellten Ergebnisse basieren auf einer Mindestverweildauer von zwei Wochen. Dieser Wert entspricht der durchschnittlichen Mindestverweildauer in der Untersuchung von Wöhner und Peters [29].

Im Unterschied zu den Verfahren von Adler et al. [1, 2] und Javanmardi et al. [11] wird in dem hier vorgestellten Ansatz die Persistenz nicht als metrische Größe interpretiert, die die exakte Zeitspanne misst, für die eine Änderung Bestand hat. Stattdessen wird eine Klassifikation in *persistente* und *nicht persistente* Änderungen vorgenommen. Im Gegensatz zu den Metriken von Javanmardi et al. [11] und Adler et al. [1] kann die Berechnung wesentlich schneller erfolgen, da deutlich weniger Textvergleiche benötigt werden. Bei Adler et al. [1] und Javanmardi et al. [11] muss bei einer gegebenen Änderung ein Abgleich mit allen folgenden Versionen durchgeführt werden, solange bis der entsprechende Textabschnitt nicht mehr vorhanden ist. Beim vorgestellten Ansatz bedarf es demgegenüber nur genau eines Vergleiches mit der entsprechenden Referenzversion. Der Ansatz von Anthony et al. [3] ist ähnlich schnell einzuschätzen, da auch hier nur ein Textvergleich erfolgt. Im Unterschied zu der hier vorgestellten Methodik berücksichtigt der Ansatz von Anthony et al. [3] jedoch keine Löschungen.

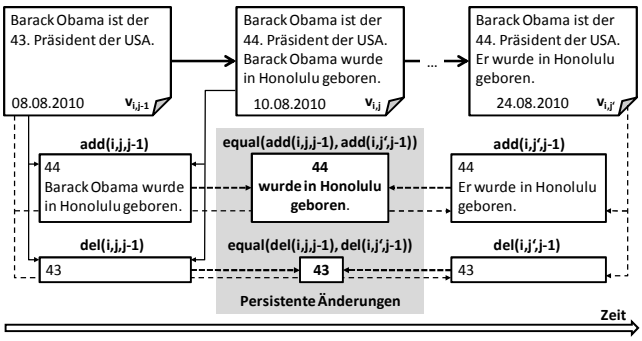

Abbildung 1: Persistente Änderungen am Beispiel

Auf Grundlage des hier vorgestellten Persistenzkriteriums wird die Anzahl der persistent geänderten Wörter eines Autors als neue potentielle Reputationsmetrik eingeführt. Die Berechnung vollzieht sich in mehreren Schritten, die in Abbildung 1 anhand eines fiktiven Beispiels veranschaulicht werden. Die Artikelversion, die nach der definierten Mindestverweildauer von zwei Wochen nach der Bearbeitung einer Artikelversion $v_{i,j}$ aktuell ist, wird im Folgenden mit $v_{i,j'}$ bezeichnet. Es werden zunächst für jede Bearbeitung des Autors sowohl die jeweils gelöschten *del(i,j,j-1)* als auch hinzugefügten Wörter *add(i,j,j-1)* berechnet. Anschließend werden die beiden Differenztexte *del(i,j',j-1)* und *add(i,j',j-1)* bestimmt, die die Änderung des Artikels innerhalb der Mindestverweildauer wiedergeben. Durch den Vergleich der berechneten Differenztexte *add(i,j,j-1)* und *add(i,j',j-1)* sowie *del(i,j,j-1)* und *del(i,j',j-1)* lässt sich ermitteln, inwieweit sich die Änderungen des Autors in den Änderungen des Artikels in der Mindestverweildauer wiederfinden und damit persistent sind. Somit lautet die Berechnungsvorschrift für die Anzahl der persistent geänderten Wörter eines Autors *a* wie folgt:

$$(4)\quad N^{pw}{}_a = \sum_i \sum_j (|equal(add(i,j',j\text{-}1), add(i,j,j\text{-}1))| + |equal(del(i,j',j\text{-}1), del(i,j,j\text{-}1))|) \quad mit\ editor(v_{i,j}) = a$$

Aufbauend auf der Persistenz $N^{pw}{}_a$ wird als weitere Metrik die Effizienz E_a eines Autors *a* vorgeschlagen. Die Effizienz E_a bezeichnet den Anteil der persistenten Änderungen $N^{pw}{}_a$ am Gesamtbearbeitungsumfang $N^{w}{}_a$ und wird mit

$$(5)\quad E_a = N^{pw}{}_a / N^{w}{}_a$$

berechnet. Geht man davon aus, dass das Verhalten eins Autors in der Vergangenheit auch Rückschlüsse auf das zukünftige Verhalten zulässt, kann die Effizienz als die Wahrscheinlichkeit interpretiert werden, mit der die Bearbeitung eines Autors von der Community akzeptiert wird.

3.2.4 Umfang der Bearbeitungen

Als weiteres Charakteristikum eines Autors wird in der Literatur der Umfang der einzelnen Bearbeitungen diskutiert [3]. Damit wird der Aufwand des Autors für seine Beiträge quantifiziert. Im Rahmen dieses Beitrages werden hierzu als Messgrößen der durchschnittliche Umfang der Bearbeitungen

$$(6)\quad Avg^{w}{}_a = N^{w}{}_a / N^{e}{}_a,$$

der durchschnittliche Umfang der persistenten Änderungen

$$(7)\quad Avg^{pw}{}_a = N^{pw}{}_a / N^{e}{}_a,$$

sowie der Umfang der größten persistenten Änderung

$$(8)\quad max^{pw}{}_a = \max_{i,j}(|equal(add(i,j',j\text{-}1), add(i,j,j\text{-}1))| + |equal(del(i,j',j\text{-}1), del(i,j,j\text{-}1))|) \quad mit\ editor(v_{i,j}) = a$$

eines Autors *a* betrachtet.

3.2.5 Beteiligung an Diskussionen

Kittur et al. [12] konnten in ihrer Studie feststellen, dass Beiträge auf den Diskussionsseiten eines Artikels zu Qualitätssteigerungen führen können. Aus diesem Grund wird mit der Anzahl der Bearbeitungen auf Diskussionsseiten

$$(9)\quad N^{diss}{}_a$$

die Beteiligung des Autors an Diskussionen als weitere Reputationsmetrik vorgeschlagen.

3.2.6 Beteiligung an qualitativ hochwertigen Artikeln

Stein und Hess [23] gehen davon aus, dass sich gute Autoren vorwiegend an qualitativ hochwertigen Artikeln beteiligen. Die Artikel, die den höchsten Qualitätsansprüchen genügen, werden in der Wikipedia als exzellente Artikel gekennzeichnet [28]. Stein und Hess verwenden daher als Reputationsmetrik das Verhältnis aus der Anzahl der Bearbeitungen an exzellenten Artikeln zur Anzahl an Bearbeitungen insgesamt. Neben den exzellenten Artikeln gelten in der Wikipedia die als lesenswert markierten Artikel ebenfalls als qualitativ hochwertig [28]. Im vorliegenden Beitrag wird deshalb in Anlehnung an Stein und Hess das Verhältnis

$$(10)\ N^{qh}{}_a = E^{qh}{}_a / N^{e}{}_a$$

aus der Anzahl der Änderungen an exzellenten und lesenswerten Artikeln $E^{qh}{}_a$ zur der Anzahl der Bearbeitungen insgesamt $N^{e}{}_a$ als weitere Reputationsmetrik vorgeschlagen.

4. EVALUATION

Die Ziele der Evaluation bestehen darin, zum einen die vorher beschriebenen Metriken vergleichend zu bewerten und deren Effizienz im Hinblick auf die Reputationsmessung zu beurteilen. Zum anderen soll eine Reputationsfunktion ermittelt werden, die die Metriken miteinander kombiniert und so eine effiziente Reputationsmessung gewährleistet. Im vorliegenden Abschnitt werden hierzu die Evaluationsmethodik und der verwendete Datenbestand vorgestellt.

Die Evaluation der vorgeschlagenen Metriken wird anhand des tatsächlichen Datenbestandes der Wikipedia durchgeführt, der als XML-Dump zur Verfügung steht. Der Datenbestand enthält sowohl die aktuellen Artikelversionen als auch die gesamte Versionsgeschichte. In der Versionsgeschichte sind die Quelltexte aller Artikelversionen inklusive HTML-Code und Wiki-Tags enthalten. Darüber hinaus beinhaltet die Versionsgeschichte zu jeder Bearbeitung weitere Metainformationen, wie den Bearbeitungszeitpunkt, den Benutzernamen des Autors und eventuelle Bearbeitungskommentare. Erfolgt die Bearbeitung durch einen nicht angemeldeten, anonymen Benutzer ist anstelle des Benutzernamens die jeweilige IP-Adresse gespeichert. Zur Evaluation wird der Datenbestand der deutschsprachigen Wikipedia vom 21. Januar 2008 ausgewertet, der zur Analyse in eine SQL-Datenbank importiert wurde. Der Datensatz enthält *646.099* enzyklopädische Artikel und *217.398* registrierte Benutzer.

Die Metriken zur automatischen Reputationsmessung werden im Folgenden anhand ihrer Trennschärfe bei der Klassifikation zwischen guten und schlechten Autoren bewertet. Dabei wird angenommen, dass eine Metrik die treffgenau gute und schlechte Autoren klassifiziert auch die Reputation der Autoren widerspiegelt. Die bei der Klassifikation erzielte Trefferquote wird daher als Maß für die Eignung einer Metrik zur Reputationsmessung interpretiert. Die Vorgehensweise, Metriken mittels einer Klassifikation zu evaluieren, ist in der Wikipediaforschung weit verbreitet und akzeptiert [5, 8, 11, 16, 24, 29, 30].

Als Grundlage für die Evaluation werden sowohl eine Menge guter als auch eine Menge schlechter Autoren benötigt, die dann anhand der Metriken zu klassifizieren sind. Hierzu wird auf Wikipedia-interne Benutzergruppierungen zurückgegriffen. Die Verwendung Wikipedia-interner Merkmale ist ebenfalls eine typische Vorgehensweise in der Wikipediaforschung. Auch bei der Evaluation von Metriken zur automatischen Qualitätsmessung werden oft Wikipedia-interne Bewertungen genutzt [5, 8, 16, 24, 29, 30].

Als Beispiele für schlechte Autoren werden gesperrte Benutzer verwendet. Dies sind Nutzer, die mutwillig Artikel zerstören oder gegen geltende Grundprinzipien der Wikipedia (Beachtung des Neutral Point of View, fairer Umgang untereinander) verstoßen und deshalb durch einen Administrator gesperrt wurden [28]. Alle anderen angemeldeten Benutzer können im Vergleich zu den gesperrten Benutzern als gute Autoren interpretiert werden. In der vorliegenden Analyse wird deshalb eine Klassifikation zwischen gesperrten Nutzern als schlechte Autoren und nicht gesperrten Benutzern als gute Autoren durchgeführt. Aus der Trefferquote dieser Klassifikation wird allgemein auf die Eignung der jeweiligen Metrik zur Reputationsmessung geschlussfolgert.

Im Unterschied zu dieser Vorgehensweise nutzen Javanmardi et al. [11] die Gruppe der Administratoren als Beispiel für gute Autoren. Administratoren sind spezielle Benutzer, die mit erweiterten Nutzerrechten ausgestattet sind und besondere Verwaltungsaufgaben wie das Sperren von Benutzern oder Wikipediaseiten und die Löschung von Seiten wahrnehmen [28]. Den Status eines Administrators erhält ein Benutzer, falls seine Kandidatur zum Administrator von der Community durch eine erfolgreiche Abstimmung bestätigt wird. Die Klassifikation zwischen Administratoren und gesperrten Benutzern ist im vorliegenden Beitrag jedoch nicht als Evaluationsmethodik geeignet. Bei der Kandidatur für den Status des Administrators wird erwartet, dass sich der entsprechende Benutzer über einen längeren Zeitraum an der Wikipedia beteiligt und mehr als *1.000* Bearbeitungen im Artikelnamensraum durchgeführt hat [28]. Somit sind bei Administratoren per Definition die Metriken der Bearbeitungshäufigkeit und des Bearbeitungsumfangs besonders intensiv ausgeprägt und das Evaluationsergebnis würde entsprechend verfälscht werden.

Die gesperrten Benutzer wurden anhand der Benutzerseiten identifiziert. Benutzerseiten sind Wikipediaseiten, auf denen sich registrierte Benutzer präsentieren können. Im Falle einer Benutzersperrung wird der ursprüngliche Inhalt von einem Administrator gelöscht und durch einen Hinweis auf die Benutzersperrung ersetzt [28]. In der Untersuchung wurden alle Benutzerseiten nach dem entsprechenden Sperrhinweis geparst.

Insgesamt enthält der Datensatz *1620* gesperrte Benutzer. Um die Gruppen der guten und schlechten Autoren in der Analyse mit gleichen Gewichten zu berücksichtigen, werden aus dem Datensatz zufällig *1620* nicht gesperrte Benutzer ausgewählt. Dabei werden alle registrierten nicht-gesperrten Benutzer mit der Ausnahme von Bots mit gleicher Wahrscheinlichkeit berücksichtigt. Insgesamt standen somit *215.642* Benutzer für die Stichprobenziehung zur Verfügung. Anonyme Benutzer werden bei der Klassifikation ausgeschlossen, da aufgrund der in der Regel dynamischen Vergabe von IP-Adressen keine dauerhaft eindeutige Zuordnung von Benutzern zu IP-Adressen gegeben ist. Die *3.240* ausgewählten Autoren sind an ca. *1,5* Millionen der insgesamt *26,3* Millionen Artikelversionen in der Wikipedia beteiligt. Damit werden durch die Analyse ca. *5%* des gesamten Artikelnamensraumes der Wikipedia ausgewertet.

Für die Autoren der Stichprobe wurden alle Messgrößen wie im vorherigen Kapitel beschrieben berechnet. Einfache Messgrößen, wie die Anzahl der Bearbeitungen $N^{e}{}_a$ oder die Anzahl der bearbeiteten Artikel $N^{p}{}_e$, lassen sich mit Hilfe einer einfachen SQL-Abfrage bestimmen. Zur Berechnung der komplexen Metriken wurde ein Satz einfacher Java-Programme implementiert.

Die Klassifikation anhand der vorgeschlagenen Metriken wird auf Basis einer linearen Diskriminanzanalyse [4] durchgeführt. Im Rahmen der Diskriminanzanalyse wird zunächst eine Diskriminanzfunktion der Form

$D(x_1...x_n) = a*x_1 + b*x_2 + \ldots + j*x_{10} + k$

berechnet.

Die Variablen x_1 bis x_{10} bezeichnen die Werte der untersuchten Metriken beim jeweiligen Autor, a bis j sind Gewichtungskoeffizienten und k ist die additive Konstante. Die Berechnung der Diskriminanzfunktion erfolgt so, dass eine bestmögliche Trennung beider Gruppen erreicht wird und sich die Funktionswerte der Diskriminanzfunktion in beiden Gruppen maximal unterscheiden.

Die Diskriminanzanalyse erfüllt damit beide einleitend genannten Evaluationsziele. Zum einen spiegeln die Gewichtungskoeffizienten die Bedeutung der einzelnen Metriken im Hinblick auf die Reputationsmessung wider. Dadurch wird der angestrebte Vergleich zwischen den verschiedenen Metriken realisiert. Zum anderen ermittelt die Diskriminanzanalyse eine effiziente Kombination der Metriken. Darüber hinaus wird im Rahmen der Diskriminanzanalyse eine Korrelationsanalyse durchgeführt und dadurch aufgedeckt, welche Metriken substituierbar sind.

In einem abschließenden Schritt wird anhand der Diskriminanzfunktion eine automatische Klassifikation zwischen nicht gesperrten und gesperrten Autoren vorgenommen und so die Güte der Diskriminanzfunktion quantifiziert. Hierzu werden zunächst die Verteilungen der Diskriminanzfunktionswerte in den beiden Gruppen berechnet. Daraus lässt sich bei gegebenem Diskriminanzfunktionswert eines Autors die Wahrscheinlichkeiten für dessen Zugehörigkeit zur Gruppe der gesperrten bzw. nicht gesperrten Autoren berechnen. Der Autor wird der Gruppe mit der höchsten Wahrscheinlichkeit zugeordnet. Die Diskriminanzanalyse wird im Rahmen der Evaluation mit Hilfe des Statistikprogramms *SPSS 17.0* durchgeführt. SPSS verwendet sowohl zur Bestimmung der Diskriminanzfunktion als auch zur Klassifikation die vollständige Stichprobe.

5. Ergebnisse

In diesem Kapitel werden die Ergebnisse der Analyse präsentiert. Zunächst werden die im Rahmen der Korrelationsanalyse aufgedeckten Abhängigkeiten der Metriken dargestellt. Danach erfolgt die vergleichende Evaluierung der Metriken. Abschließend wird die Diskriminanzfunktion als effektive Kombination der Metriken vorgestellt und bewertet.

5.1 Korrelationsanalyse

Die im Rahmen der Diskriminanzanalyse berechnete Korrelationsmatrix ist in Tabelle 2 dargestellt. Mit einem Korrelationskoeffizienten von *0,983* korrelieren die Anzahl der Bearbeitungen N^e_a und die Anzahl der bearbeiteten Artikel N^p_a sehr stark. Dies bedeutet, dass Autoren die viele Bearbeitungen durchführen auch viele verschiedene Artikel bearbeiten. Des Weiteren zeigt sich eine sehr starke Korrelation von *0,999* zwischen dem Gesamtbearbeitungsumfang N^w_a und der Anzahl der persistent geänderten Wörter N^{pw}_a.

Aufgrund der hohen Korrelationen zwischen N^e_a und N^p_a sowie zwischen N^w_a und N^{pw}_a sind die Variablen eines Paares miteinander substituierbar. Es ist daher sinnvoll, für die weitere Analyse jeweils nur eine Metrik stellvertretend für jedes Variablenpaar zu betrachten. Für die Bestimmung der Diskriminanzfunktion werden N^e_a und N^w_a gewählt, da deren Berechnung weniger aufwändig ist.

Häufig wird gegenüber der Wikipedia die Kritik geäußert, dass insbesondere Änderungen von erfahrenen Autoren von der Community akzeptiert werden. Diese Kritik kann durch die vorliegende Untersuchung widerlegt werden. Messgrößen, die die Erfahrung eines Autors quantifizieren (N^p_a, N^e_a, N^w_a, N^{pw}_a), korrelieren mit Korrelationskoeffizienten zwischen *0,044* bis *0,055* nur sehr gering mit der Effizienz E_a.

Tabelle 2: Korrelation der Metriken

Korrelation	N^e_a	N^p_a	N^w_a	N^{pw}_a	max^{pw}_a
N^e_a	1	0,983	0,561	0,559	0,321
N^p_a	0,983	1	0,517	0,517	0,292
N^w_a	0,561	0,517	1	0,999	0,719
N^{pw}_a	0,559	0,517	0,999	1	0,719
max^{pw}_a	0,321	0,292	0,719	0,719	1
E_a	0,052	0,054	0,044	0,047	0,039
Avg^w_a	0	0	0,012	0,008	0,008
Avg^{pw}_a	-0,01	-0,02	0,146	0,145	0,135
N^{qh}_a	-0,06	-0,07	-0,041	-0,04	-0,027
N^{diss}_a	0,434	0,38	0,383	0,371	0,198
Korrelation	E_a	Avg^w_a	Avg^{pw}_a	N^{qh}_a	N^{diss}_a
N^e_a	0,052	-0,004	-0,007	-0,062	0,434
N^p_a	0,054	-0,004	-0,017	-0,066	0,38
N^w_a	0,044	0,012	0,146	-0,041	0,383
N^{pw}_a	0,047	0,008	0,145	-0,041	0,371
max^{pw}_a	0,039	0,008	0,135	-0,027	0,198
E_a	1	-0,129	0,186	-0,129	0,053
Avg^w_a	-0,129	1	0,21	0,188	-0,005
Avg^{pw}_a	0,186	0,21	1	0,063	0,055
N^{qh}_a	-0,129	0,188	0,063	1	-0,043
N^{diss}_a	0,053	-0,005	0,055	-0,043	1

5.2 Vergleich der Metriken

In der Diskriminanzanalyse beschreibt die Strukturmatrix die Bedeutung der verwendeten Metriken bei der Klassifikation. Die Koeffizienten der Strukturmatrix geben die Korrelation der jeweiligen Metrik zur Diskriminanzfunktion an. Dementsprechend lässt sich auf Basis der Strukturmatrix die Aussagekraft der verwendeten Metriken für die Reputationsmessung vergleichend bewerten. Die Strukturmatrix ist in Tabelle 3 dargestellt.

Die Analyse zeigt, dass sich gute und schlechte Autoren besonders hinsichtlich der Effizienz E_a unterscheiden. Mit *0,965* erhält diese Metrik ein sehr hohes Gewicht in der Strukturmatrix. Im Vergleich zu allen anderen Metriken kann damit die Reputation eines Autors am besten beschrieben werden. In Tabelle 4 sind die Mittelwerte μ und die Standardabweichungen σ der untersuchten Metriken in den beiden Autorengruppen aufgelistet. Die Aussagekraft von E_a lässt sich auch an den entsprechenden Mittelwerten erkennen. So haben gesperrte Nutzer eine durchschnittliche Effizienz E_a von *23%* und nicht gesperrte Benutzer von *85%*. Die Analyse belegt daher, dass die Änderungen von schlechten Autoren zum größten Teil durch die Community verworfen werden,

wohingegen die Änderungen von guten Autoren mit einer hohen Wahrscheinlichkeit akzeptiert werden. Die Studien von Adler et al. [1], Javanmardi et al. [11] und Anthony et al. [3], die als Reputationsmetrik eine zu E_a ähnliche Metrik vorschlagen, werden somit durch die vorliegende Analyse bestätigt.

Tabelle 3: Vergleich der Metriken

	Korrelation zur Diskriminanzfunktion (Strukturmatrix)
$\mathbf{E_a}$	0,965
$\mathbf{Avg^{pw}_a}$	0,2
$\mathbf{N^{diss}_a}$	0,162
$\mathbf{N^{w}_a}$	0,147
$\mathbf{N^{e}_a}$	0,141
$\mathbf{N^{qh}_a}$	0,112
$\mathbf{Avg^{w}_a}$	-0,111
$\mathbf{max^{pw}_a}$	0,109

Mit Ausnahme der Effizienz E_a korrelieren alle anderen Metriken mit der Diskriminanzfunktion ähnlich stark. Die jeweiligen Korrelationskoeffizienten liegen im Bereich von *0,109* bis *0,2*. Damit kann neben der Effizienz allen weiteren untersuchten Metriken eine ähnliche Aussagekraft zugeschrieben werden. Diese ist jedoch im Vergleich zur Effizienz deutlich geringer. Die Mittelwerte der Metriken in den beiden Autorengruppen zeigen, dass alle Metriken bei den nicht gesperrten Nutzern höhere Werte als bei den gesperrten Nutzern aufweisen. Beispielsweise führen nicht gesperrte Nutzer im Durchschnitt persistente Änderungen (N^{pw}_a) im Umfang von ca. *82.000* Wörtern und gesperrte Benutzer im Durchschnitt von nur ca. *1.600* Wörtern durch.

Tabelle 4: Mittelwerte und Standardabweichungen bei gesperrten und nicht gesperrten Benutzern

	gesperrte Benutzer		nicht gesperrte Benutzer	
	μ	σ	μ	σ
$\mathbf{N^{e}_a}$	65	598	2.774	12.629
$\mathbf{N^{p}_a}$	29	295	1.796	8.473
$\mathbf{N^{w}_a}$	3.116	22.476	88.346	419.374
$\mathbf{N^{pw}_a}$	1.631	18.262	81.988	403.512
$\mathbf{max^{pw}_a}$	167	942	3.705	22.506
$\mathbf{E_a}$	23,00%	33,00%	85,00%	19,00%
$\mathbf{Avg^{w}_a}$	219	904	44	159
$\mathbf{Avg^{pw}_a}$	11	47	32	43
$\mathbf{N^{qh}_a}$	7,00%	20,00%	13,00%	21,00%
$\mathbf{N^{diss}_a}$	16	109	128	454

Als einzige Ausnahme zeigt sich beim durchschnittlichen Umfang der Bearbeitungen Avg^{w}_a ein umgekehrter Trend. Hier deutet ein kleiner Wert auf eine gute Reputation hin. Während nicht gesperrte Benutzer durchschnittlich *44* Wörter pro Bearbeitung ändern, ändern gesperrte Benutzer im Durchschnitt *219* Wörter. Bei der durchschnittlichen persistenten Änderung pro Bearbeitung Avg^{pw}_a ist diese Tendenz nicht zu beobachten. Dies lässt sich damit erklären, dass gesperrte Benutzer umfangreiche Bearbeitungen wie das Löschen des gesamten Textes oder das Hinzufügen langer, unsinniger Textabschnitte vornehmen und die entsprechenden Bearbeitungen sehr schnell von anderen Wikipedia-Benutzern korrigiert werden.

Zusätzlich zur Strukturmatrix lässt sich ein Vergleich der Metriken realisieren, indem jede einzelne Metrik im Rahmen einer Diskriminanzanalyse isoliert betrachtet wird. Die Trefferquoten dieser Auswertung sind in Tabelle 5 abgebildet.

Tabelle 5: Trefferquoten der Einzelmetriken

	Trefferquote
$\mathbf{E_a}$	85,1%
$\mathbf{Avg^{pw}_a}$	67,3%
$\mathbf{N^{diss}_a}$	59,9%
$\mathbf{N^{w}_a}$	61,2%
$\mathbf{N^{e}_a}$	62,3%
$\mathbf{N^{qh}_a}$	58,4%
$\mathbf{Avg^{w}_a}$	56,8%
$\mathbf{max^{pw}_a}$	64,7%

Diese Analyse bestätigt die besondere Bedeutung der Effizienz E_a, die sich bereits aus der Strukturmatrix ergibt. Mit der Effizienz konnte die deutlich höchste Trefferquote von *85,1%* erzielt werden. Trotz der deutlichen Unterschiede der Mittelwerte μ zwischen gesperrten und nicht gesperrten Autoren (Tabelle 4) erreichen alle anderen Metriken eine vergleichsweise niedrige Trefferquote zwischen *56,8%* und *67,3%*. Dies deutet darauf hin, dass die deutlichen Unterschiede der Mittelwerte auf wenige Autoren mit besonders extremen Werteausprägungen zurückzuführen sind. Diese Auswertung belegt, dass sich die Metriken nicht für eine isolierte Verwendung in einem Reputationssystem eignen, da die Reputation ungenau erfasst wird. Diese sind Metriken hauptsächlich für eine kombinierte Reputationsfunktion anwendbar, um das Bearbeitungsverhalten der Autoren möglichst umfassend abzubilden.

5.3 Definition und Bewertung der Diskriminanzfunktion

Die Diskriminanzanalyse fügt die verschiedenen Metriken zu einer gemeinsamen Funktion zusammen. Die Gewichtung der einzelnen Metriken in der Diskriminanzfunktion entspricht dabei ihrer Signifikanz für die Klassifikation zwischen guten und schlechten Autoren.

Tabelle 6: Diskriminanzfunktion

	Gewichtungskoeffizient
$\mathbf{E_a}$	3,63979532
$\mathbf{Avg^{pw}_a}$	-0,00009770
$\mathbf{N^{diss}_a}$	0,00020281
$\mathbf{N^{w}_a}$	0,00000012
$\mathbf{N^{e}_a}$	0,00000437
$\mathbf{N^{qh}_a}$	1,15645448
$\mathbf{Avg^{w}_a}$	-0,00004466
$\mathbf{max^{pw}_a}$	0,00000104
Konstante	-2,12285923

Die von SPSS berechnete effiziente Diskriminanzfunktion ist in Tabelle 6 dargestellt. Aufgrund der bestehenden Korrelationen werden für einige Metriken auch negative Gewichtungskoeffizienten berechnet. Die ermittelten Koeffizienten sind nicht ausschließlich von der Aussagekraft der Metrik abhängig, sondern werden ebenfalls durch den jeweiligen Wertebereich beeinflusst. Da die Effizienz $(0 \leq E_a \leq 1)$ und die Beteiligung bei qualitativ hochwertigen Artikeln $(0 \leq N^{qh}{}_a \leq 1)$ im Vergleich zu den anderen Metriken sehr kleine Wertebereiche aufweisen (Tabelle 4), werden beide Metriken mit sehr hohen Gewichten berücksichtigt.

Anhand der Klassifikationsgüte lässt sich die Güte der Diskriminanzanalyse beurteilen. Tabelle 7 stellt das Ergebnis der Klassifikation auf Basis der Diskriminanzanalyse dar. Durch die berechnete Diskriminanzfunktion werden insgesamt *86,5%* der Autoren des Datensatzes richtig klassifiziert. Die True-Positive-Rate (TPR) bezeichnet den Anteil der richtig klassifizierten nicht gesperrten Benutzer und beträgt *93,5%.* Die False-Positive-Rate (FPR) als Anteil der fälschlicherweise als nicht gesperrt klassifizierten gesperrten Benutzer beträgt *20,6%*. Die hohen Trefferquoten zeigen, dass sich die Diskriminanzfunktion gut zur Reputationsmessung in der Wikipedia eignet.

Tabelle 7: Klassifikationsergebnis

	Anzahl	**Anteil**
TPR	1.514	93,50%
FPR	333	20,60%
Trefferquote	***2.801***	***86,50%***

Zur Analyse der Fehlklassifikationen wurden strichprobenartig einige falsch klassifizierte Testfälle ausgewählt und deren Änderungen in der Wikipedia manuell nachvollzogen. Bei gesperrten Nutzern wurde ergänzend das *Benutzersperr-Logbuch* ausgewertet. Darin sind alle Sperrvorgänge und die durch die Administratoren benannten Gründe für die Sperrung aufgelistet.

Die Fehlklassifikationen bei den gesperrten Nutzern lassen sich größtenteils darauf zurückführen, dass die Sperrung aufgrund von Meinungsverschiedenheiten bzw. Konflikten mit anderen Nutzern oder Administratoren ausgesprochen wurde. Die jeweiligen Autoren haben in der Regel vor der Sperrung umfangreiche Bearbeitungen mit einer hohen Effizienz vorgenommen, so dass ein entsprechend hoher Diskriminanzfunktionswert berechnet wird. Fehler bei den nicht gesperrten Benutzern sind darauf zurückzuführen, dass sich innerhalb dieser Gruppe auch schlechte Benutzer befinden, die beispielsweise Vandalismus begehen.

Durch die Kombination der verschiedenen Metriken in der Diskriminanzfunktion kann die Effektivität der Reputationsmessung gegenüber einer Einzelbetrachtung der Metriken verbessert werden. Im Vergleich zur Einzelbetrachtung der Effizienz (Tabelle 5) kann die Trefferquote um *1,4%* verbessert werden. Bei den anderen Metriken fällt dieser Unterschied wesentlich deutlicher aus.

Ein weiterer Vorteil der hier berechneten Diskriminanzfunktion zeigt sich am Beispiel der Administratoren. Diese als sehr gut geltenden Autoren erzielen die höchsten Diskriminanzfunktionswerte. Durch die Kombination der Metriken werden verschiedene Facetten ihres Bearbeitungsverhaltens abgebildet, sodass sich Administratoren hinsichtlich ihrer Reputation deutlicher von anderen Benutzern abgrenzen können.

6. SCHLUSSBETRACHTUNG

Im Rahmen der Schlussbetrachtung wird zunächst der Beitrag kurz zusammengefasst, anschließend die Untersuchung kritisch gewürdigt und ein Ausblick auf weitere Forschungsfragen gegeben.

6.1 Zusammenfassung

In diesem Beitrag wurden zunächst potentielle Messgrößen zur automatischen Messung der Autorenreputation vorgestellt. Die Metriken umfassen die Kategorien Bearbeitungshäufigkeit, Gesamtbearbeitungsumfang, Umfang der einzelnen Bearbeitungen, Beteiligung an qualitativ hochwertigen Artikeln, Beteiligung an Diskussionen und die Persistenz der Bearbeitungen. Die Metriken wurden aus der bestehenden Literatur entnommen und um eigene Vorschläge ergänzt. In der Kategorie der Persistenz wurde ein neuer Berechnungsansatz vorgestellt. Während bestehende Ansätze genau die Zeitspanne bestimmen, die eine Änderung Bestand hat, werden in diesem Beitrag Änderungen dann als persistent eingestuft, wenn sie für mindestens zwei Wochen Bestand haben. Dadurch wird die Anzahl der notwendigen Textvergleiche deutlich reduziert und die Berechnung kann wesentlich schneller erfolgen. Dies ist ein wichtiger Aspekt im Hinblick auf eine Implementierung in der Praxis, da seitens der Wikipedia insbesondere Performanceprobleme als Hürde bei der Einführung neuer Konzepte genannt werden.

Des Weiteren wurden erstmalig in der Wikipediaforschung zur Reputationsmessung verschiedene Metriken anhand einer Diskriminanzanalyse vergleichend evaluiert. Die neu vorgeschlagene Metrik der *Effizienz der Bearbeitungen* erhält dabei ein sehr hohes Gewicht und eignet sich demnach besonders gut zur Reputationsmessung. Die vorliegende Studie bestätigt damit die Forschungsarbeiten von Adler et al. [1, 2], Javanmardi et al. [11] sowie Anthony et al. [3], die eine vergleichbare Metrik vorgeschlagen haben.

Als weiteres Ergebnis wurden erstmals verschiedene Metriken durch die Diskriminanzanalyse zu einer Reputationsfunktion zusammengeführt. Bei der Gestaltung einer automatischen Reputationsmessung besteht ein Aspekt darin, das Bearbeitungsverhalten der Autoren möglichst vollständig zu erfassen. Durch die vorgeschlagene Kombination der Metriken können verschiedene Facetten des Bearbeitungsverhaltens der Autoren quantifiziert werden. Deshalb kann im Vergleich zur isolierten Betrachtung der Metriken, welche in den bestehenden Forschungsarbeiten angewendet wird, eine aussagekräftigere Reputationsmessung erreicht werden.

Anhand der berechneten Reputationsfunktion wird mit einer Trefferquote von *86,5%* zwischen guten (nicht gesperrten Benutzern) und schlechten Autoren (gesperrten Benutzern) unterschieden. Vor allem gute Autoren konnten mit einer True-Positive-Rate von *93,5%* sehr gut klassifiziert werden. Die Trefferquoten zeigen, dass die ermittelte Reputationsfunktion eine effektive automatische Reputationsmessung in der Wikipedia ermöglicht.

6.2 Kritische Würdigung und Ausblick

Die vorgestellte Studie zeigt, dass in der Wikipedia eine automatische Klassifikation von Benutzergruppen effizient umsetzbar ist. Diese Idee könnte zukünftig genutzt werden, um beispielsweise den Status des Administrators automatisch zu vergeben. Eine solche Vorgehensweise wird in der deutschen Wikipedia bereits beim Sichterstatus angewendet. Durch eine Diskriminanzanalyse zwischen Administratoren und angemeldeten nicht-gesperrten

Benutzern kann eine effektive Kombination von Metriken zur Vergabe des Administratorstatus berechnet werden.

Die hier eingeführte Metrik der Effizienz der Bearbeitung kann in weiteren Forschungsarbeiten mit den Ansätzen von Adler et. al. [1, 2] und Javanmardi et. al. [11] verglichen werden. Durch eine Implementierung dieser beiden Ansätze kann der Performancegewinn und die Aussagekraft im Vergleich zu den hier vorgestellten Verfahren quantifiziert werden.

Ferner kann untersucht werden, inwieweit sich ich das vorgeschlagene Reputationssystem im Hinblick auf die Effizienz der Reputationsmessung verbessern lässt. Hierzu können zunächst die Autoren anhand der Metriken geclustert werden, um so typische Bearbeitungsmuster aufzudecken. Beispielsweise ist es denkbar, dass einige Benutzer vorwiegend Korrekturen vornehmen oder Vandalismus beseitigen. Andere Benutzer wiederum schreiben die Artikel fort und fügen neuen Text hinzu. Für die identifizierten Cluster lassen sich möglicherweise spezielle Reputationsfunktionen definieren, die zu einer effizienteren Reputationsmessung führen. Die clusterbasierte Reputationsmessung kann beispielsweise mit Hilfe von Entscheidungsbäumen umgesetzt werden.

Des Weiteren soll das Reputationssystem prototypisch in die von der Wikipedia verwendete MediaWiki-Software integriert werden. In praktischen Tests sollen dann anhand des Prototyps die Wirkung auf das Benutzerverhalten und die Aussagekraft des Reputationssystems experimentell validiert werden. Darüber hinaus bietet ein Prototyp die Möglichkeit, eventuelle Performance-Einbußen aus der Reputationsberechnung abzuschätzen.

Das hier vorgestellte Reputationssystem kann in der Wikipedia in verschiedener Weise genutzt werden. So bietet das Reputationssystem eine Basis für ein effektives Rechtemanagement. Auch motiviert das Reputationssystem Autoren sich intensiv an der Wikipedia zu beteiligen. Eine weitere Anwendung besteht darin, anhand der Autorenreputation einen Qualitätsscore für die Artikel zu berechnen. Ein solches Verfahren wird in einer folgenden Publikation vorgestellt. Erste Untersuchungen zeigen, dass dadurch eine sehr effiziente Qualitätsmessung möglich ist.

7. LITERATUR

[1] Adler, B.T. und Alfaro, L. 2007. A Content-Driven Reputation System for the Wikipedia. In *Proceedings of the 16th International Conference on the World Wide Web* (Banff, Kanada, 08.-12. Mai 2007). WWW2007. ACM, New York, NY, 261-270. DOI=http://doi.acm.org/10.1145/1242572.1242608.

[2] Adler, B.T., Chatterjee, K., de Alfaro, L., Faella, M., Pye, I. und Raman, V. 2008. Assigning Trust To Wikipedia Content. In *Proceedings of the 2008 International Symposium on Wikis* (Porto, Portugal, 08.-10. September 2008). WikiSym 2008. ACM, New York, NY. DOI=http://doi.acm.org/10.1145/1822258.1822293.

[3] Anthony, D., Smith, S. W. und Williamson, T. 2007. *The Quality of Open Source Production: Zealots and Good Samaritans in the Case of Wikipedia*. Technical Report, TR2007-606. Department of Computer Science, Dartmouth College.

[4] Backhaus, K., Erichson, B., Wulff, P. und Weiber, R. 2008. *Multivariante Analysemethoden – Eine anwendungsorientierte Einführung*. Springer, Berlin und Heidelberg.

[5] Blumenstock, J.E. 2008. Size Matters: Word Count as a Measure of Quality on Wikipedia. In *Proceedings of the 17th international conference on World Wide Web* (Peking, China, 21.-25. April 2008). WWW08. ACM, New York, NY, 1095-1096. DOI=http://doi.acm.org/10.1145/1367497.1367673.

[6] Cunningham, W. und Leuf, B. 2001. *The Wiki Way. Quick Collaboration on the Web*. Addison-Wesley, Boston u.a.

[7] Denning, P., Horning, J., Parnas, D. und Weinstein, L.. 2005. Wikipedia risks. *Communications of the ACM*. 48, 12 (Dezember 2005), 152-152. DOI=http://doi.acm.org/10.1145/1101779.1101804.

[8] Dondio, P. und Barrett, S. 2007. Computational Trust in Web Content Quality: A Comparative Evalutation on the Wikipedia Project. *Informatica – An International Journal of Computing and Informatics*. 31, 2 (Juni 2007), 151-160. DOI=10.1.1.159.4166.

[9] Giles, G. 2005. Internet encyclopedias go head to head. *Nature*, 438, 7070 (Dezember 2005), 900-901. DOI=10.1038/438900a.

[10] Hunt, J. und McIlroy, M. 1975. *An algorithm for differential file comparison*. Computer Science Technical Report 41, Bell Laboratories.

[11] Javanmardi, S., Lopes, C. und Baldi, P. 2010. Modeling User Reputation in Wikipedia. *Journal of Statistical Analysis and Data Mining*. 3, 2 (April 2010), 126-139. DOI=10.1002/sam.10070.

[12] Kittur, A. und Kraut, R. E. 2008. Harnessing the Wisdom of Crowds in Wikipedia: Quality Through Coordination. In *Proceedings of the ACM 2008 Conference on Computer supported cooperative work* (San Diego, USA, 8.-12. November 2008). CSCW08. ACM, New York, NY, 37-46. DOI=http://doi.acm.org/10.1145/1460563.1460572.

[13] Kurzidim, M. 2004. Wissenswettstreit. Die kostenlose Wikipedia tritt gegen die Marktführer Encarta und Brockhaus an. *c't Magazin für Computertechnik*, 2004, 21 (Oktober 2004), 132-139.

[14] Lih, A. 2004. Wikipedia as participatory journalism: Reliable sources? Metrics for evaluating collaborative media as a news resource. In *Proceedings of the 5th International Symposium on Online Journalism* (Austin, USA, 16.-17. April 2004). DOI=10.1.1.117.9104.

[15] Lim, E.P., Vuong, B.Q., Lauw, H.W. und Sun, A. 2006. Measuring Qualities of Articles Contributed by Online Communities. In *Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence* (Hong Kong, 18.-22. Dezember 2006). WI '06, IEEE Computer Society, Washington, DC, 81-87. DOI=10.1109/WI.2006.115.

[16] Opuszko, M., Wöhner, T., Peters, R. und Ruhland, J., 2010. Qualitätsmessung in der Wikipedia: Ein Ansatz auf Basis von Markov-Modellen. In *Multikonferenz Wirtschaftsinformatik 2010* (Göttingen, Deutschland, 23.-25. Februar 2010). MKWI 2010, Universitätsverlag Göttingen, 705-716.

[17] O'Reilly, T. 2005. *What is Web2.0?* http://oreilly.com/web2/archive/what-is-web-20.html. Abruf am 02.12.2010.

[18] Potthast, M., Stein, B. und Gerling, R. 2008. Automatic Vandalism Detection in Wikipedia. In *Proceedings of the Advances in Information Retrieval - 30th European Conference on IR Research*. (Glasgow, UK, 30. März-3. April 2008). ECIR 2008, Springer, 663-668. DOI=10.1007/978-3-540-78646-7_75.

[19] Priedhorsky, R., Chen, J., Lam, S.K., Panciera, K., Terveen, L. und Riedl, J. 2007. Creating, Destroying, and Restoring Value in Wikipedia. In *Proceedings of the 2007 international ACM conference on Supporting group work* (Sanibel Island, USA, 4.-7. November 2007). Group 2007, ACM, New York, NY, 259-268. DOI=http://doi.acm.org/10.1145/1316624.1316663.

[20] Resnick, P., Zeckhauser, R., Friedman, E. und Kuwabara, K. 2000. Reputation Systems. *Communications of the ACM*, 43, 12 (Dezember 2000). 45-48. DOI=http://doi.acm.org/10.1145/355112.355122.

[21] Smets, K., Goethals, B. und Verdonk, B. 2008. Automatic Vandalism Detection in Wikipedia: Towards a Machine Learning Approach. In *Proceedings of the AAAI Workshop, Wikipedia and Artificial Intelligence: An Evolving Synergy* (Chicago, USA, 13.-14. Juli 2008), WikiAI 2008, 43-48.

[22] Surowiecki, J.; 2004. *The Wisdom Of Crowds: Why The Many Are Smarter Than The Few And How Collective Wisdom Shapes Business, Economies, Societies And Nations*. Doubleday.

[23] Stein, K. und Hess, C. 2007. Does it matter who contributes: a study on featured articles in the german Wikipedia. In *Proceedings of the eighteenth conference on Hypertext and hypermedia* (Manchester, UK, 10.-12. September 2007). HT '07, ACM, New York, NY, 171-174. DOI=http://doi.acm.org/10.1145/1286240.1286290:

[24] Stvilia, B., Twidale, M.B., Smith, L.C. und Gasser, L. 2005. Assessing information quality of a community-based encyclopedia. In *Proceedings of the International Conference on Information Quality* (Cambridge, USA, 4.-6. November 2005). ICIQ 2005, 442–454. DOI=*10.1.1.78.6243*.

[25] Viégas, F., Wattenberg, M. und Dave, K. 2004. Studying cooperation and conflict between authors with history flow visualizations. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Wien, Österreich, 24.-29. April 2004). CHI 2004, ACM, New York, NY, 575–582. DOI=http://doi.acm.org/10.1145/985692.985765.

[26] Viégas, F., Wattenberg, M., Kriss, J. und Ham, F. 2007. Talk before you type: Coordination in Wikipedia. In *Proceedings of the 40th Hawaii International Conference on System Sciences*. (Hawaii, USA, 3.-6. Januar 2007). HICSS 2007, IEEE Computer Society Washington, DC, 78-88. DOI=http://dx.doi.org/10.1109/HICSS.2007.511.

[27] West, A., Kannany, S. und Leez, I. 2010. *Detecting Wikipedia Vandalism via Spatio-Temporal Analysis of Revision Metadata*. Technical Reports (CIS). University of Pennsylvania.

[28] Wikipedia (Hrsg.). 2010. *Autorenportal*. http://de.wikipedia.org/wiki/Wikipedia:Autorenportal. Abruf am 02.12.2010.

[29] Wöhner, T. und Peters, R. 2009. Assessing the quality of Wikipedia articles with lifecycle based metrics. In *Proceedings of the 5th International Symposium on Wikis and Open Collaboration* (Orlando, USA, 25.-27. Oktober 2009). WikiSym 2009, ACM, New York, NY. DOI=http://doi.acm.org/10.1145/1641309.1641333.

[30] Zeng, H., Alhoussaini, M., Ding, L., Fikes R.und McGuinness, D. 2006. Computing trust from revision history. In *Proceedings of the 2006 International Conference on Privacy, Security and Trust* (Markham, Kanada, 30. Oktober-1. November 2006). PST 2006, ACM, New York, NY. DOI=http://doi.acm.org/10.1145/1501434.1501445.