Qualitätsmessung in der Wikipedia

Ein Ansatz auf Basis von Markov-Modellen

Marek Opuszko[1], Thomas Wöhner[2], Ralf Peters[2], Johannes Ruhland[1]

[1]Lehrstuhl für Wirtschaftsinformatik, Friedrich-Schiller-Universität Jena
Carl-Zeiss-Str. 3, 07743 Jena
[2]Lehrstuhl für Wirtschaftsinformatik, insb. E-Business,
Martin-Luther Universität Halle-Wittenberg
Universitätsring 3 , 06108 Halle (Saale)

1 Einleitung

Charakteristisch für Web2.0-Anwendungen ist die kooperative Inhaltserstellung durch die Internetnutzer selbst. Nach dem Zerplatzen der „Dotcom-Blase" im Jahr 2001 gewannen solche Anwendungen im WWW zunehmend an Bedeutung (O'Reilly 2005). Ein besonders verbreiteter Anwendungstyp innerhalb des Web2.0 sind Wikis. Es handelt sich dabei um Websites, deren Inhalte zumeist ohne vorherige Anmeldung und ohne besondere Vorkenntnisse direkt im Webbrowser bearbeitet werden können (Leuf und Cunningham 2004, S. 14; Hippner und Wilde 2005, S. 442). Als bedeutendste Wiki gilt die freie Online-Enzyklopädie Wikipedia, die weltweit mehr als 13.000.000 Artikel in mehr als 260 Sprachen enthält. Die größte Wikipedia ist die englische Ausgabe mit mehr als 3.000.000 Artikeln, gefolgt von der deutschen Ausgabe mit ca. 950.000 Artikeln (Zachte 2009).[1]

Der offene Zugang der Wikipedia führt zwar zu einer hohen Beteiligung, sodass in verhältnismäßig kurzer Zeit qualitativ hochwertige Artikel entstehen und zeitnah aktualisiert werden. Jedoch birgt diese Offenheit auch Risiken. So ist es beispielsweise nicht ausgeschlossen, dass Artikel mutwillig manipuliert werden oder aufgrund mangelnder Expertise der Autoren fehlerhafte Informationen enthalten (Denning et al. 2005, S. 152). Im Ergebnis können Benutzer die Qualität eines Wikipedia-Artikels nur schwer einschätzen. Zur Lösung dieses Problems wurde in der Wikipedia eine benutzergetriebene Qualitätsbewertung eingeführt. So können qualitativ hochwertige Artikel durch ein abstimmungsbasiertes Verfahren als *Lesenswert* oder *Exzellent* gekennzeichnet werden. Qualitativ minderwertige Artikel können als *Löschkandidat* markiert werden (Wikipedia 2009). Durch diese Wikipedia-internen Verfahren wird allerdings nur ein kleiner Teil der Wikipedia-

[1] Stand Juli 2009

Artikel erfasst.[2] Darüber hinaus zeichnet sich die Wikipedia durch eine hohe Dynamik aus, sodass getätigte Bewertungen schnell obsolet werden können. Aufgrund dieser Schwächen der benutzergetriebenen Qualitätsbewertungen gewinnt die automatische Qualitätsmessung zunehmend an Interesse und hat sich in den letzten Jahren als eigenständiges Forschungsgebiet etabliert.

Einen interessanten Ansatzpunkt zur automatischen Qualitätsbewertung bietet der Lebenszyklus eines Artikels (Wöhner und Peters 2009). Hierunter wird die Entwicklung der Bearbeitungsintensität über die gesamte Lebensspanne eines Artikels hinweg verstanden. Der Lebenszyklus lässt sich somit prinzipiell als Zeitreihe modellieren. Gegenwärtig fehlt es jedoch in der Forschung an geeigneten Verfahren zur Zeitreihenanalyse, um Lebenszyklen detailliert analysieren zu können. In diesem Beitrag wird deshalb ein neues Verfahren zur Zeitreihenanalyse auf Basis von Markov-Modellen vorgestellt. Es wird gezeigt, dass durch dieses Verfahren bestehende Ansätze zur automatischen Qualitätsmessung in der Wikipedia sinnvoll ergänzt und verbessert werden können.

2 Stand der Forschung

2.1 Automatische Qualitätsmessung in der Wikipedia

Hinsichtlich der automatischen Qualitätsmessung in der Wikipedia kann zwischen Verfahren, die die Vertrauenswürdigkeit einzelner Wörter oder Textabschnitte bewerten, und Verfahren, die die Qualität eines gesamten Artikels messen, unterschieden werden. Die Verfahren der ersten Kategorie (Cross 2006; Adler und de Alfaro 2007; S. 261ff.; Adler et al. 2008) ermöglichen es, die Richtigkeit einzelner Fakten eines Artikels abzuschätzen. Anhand der Verweildauer der Wörter bzw. Textabschnitte wird auf deren Vertrauenswürdigkeit geschlossen und dies durch unterschiedliche Einfärbung des Textes signalisiert.

Der größere Teil der Forschungsarbeiten zur automatischen Qualitätsmessung befasst sich jedoch mit Verfahren zur Bewertung gesamter Artikel (Lih 2004; Lim et al. 2006, S. 81ff.; Zeng et al. 2006; Stvilia et al. 2005, S. 442ff.; Dondio und Barrett 2007, S. 151ff. sowie Blumenstock 2008, S. 1095f.). Die Verfahren verwenden dabei größtenteils einfache Kriterien wie die Artikellänge, die Anzahl der Autoren oder die Anzahl der Versionen. Da in den genannten Publikationen ähnliche Evaluationsmethodiken genutzt werden, sind die Ergebnisse vergleichbar. So konnte Blumenstock (2008) zeigen, dass die Artikellänge das bis dahin effektivste Kriterium zur Qualitätsmessung ist.

[2] So enthielten im Januar 2008 nur ca. 3500 von insgesamt ca. 650.000 Artikeln der deutschsprachigen Wikipedia eine solche Bewertung.

2.2 Lebenszyklus von Wikipedia-Artikeln

Der Lebenszyklus eines Artikels wird von Wöhner und Peters (2009) anhand der Entwicklung der Bearbeitungsintensität über die gesamte Lebensspanne untersucht. Die Bearbeitungsintensität wird dabei anhand der persistenten und der transienten Änderungen gemessen. Die persistenten Änderungen fassen alle Bearbeitungen zusammen, die den Artikel effektiv ändern. Transiente Änderungen hingegen sind flüchtige Bearbeitungen, die schnell verworfen werden und damit nicht zur Fortschreibung des Artikels beitragen. Da Bearbeitungen wie Vandalismus oder Ähnliches nach sehr kurzer Zeit (in der Regel innerhalb von 3 Minuten) durch die Wikipedia-Community korrigiert werden (Víegas et al. 2004, S. 579), sind diese Änderungen insbesondere durch die transienten Änderungen repräsentiert. Beide Messgrößen werden monatsweise berechnet. Der Umfang wird durch die Anzahl an Wörtern definiert, die durch die jeweiligen Bearbeitungen gelöscht bzw. neu eingefügt wurden. Der Vergleich von qualitativ minderwertigen und hochwertigen Artikeln erfolgt anhand einer einfachen Durchschnittsbildung. Dabei zeigen beide Artikelkategorien charakteristische Eigenschaften, mit denen sich eine im Vergleich zu den bestehenden Verfahren verbesserte Qualitätsmessung erreichen lässt.

Die Aggregation der Lebenszyklen mittels einer einfachen Durchschnittsbildung lässt jedoch verschiedene Aspekte des zeitlichen Verlaufes unberücksichtigt. So werden Artikel, die besonders intensiv bearbeitet wurden, stärker gewichtet. Auch variiert aufgrund der unterschiedlichen Entstehungsdaten und Bewertungszeitpunkte der Artikel die Lebensspanne, sodass es bei der Durchschnittsbildung zu Verzerrungen kommt. Innerhalb einer Qualitätskategorie sind zudem in jeder Periode Messwerte vorhanden, sodass sich im Durchschnitt ein kontinuierlicher Verlauf der Messgrößen ergibt. Der Lebenszyklus eines Einzelartikels ist aber in der Regel durch einen stark schwankenden Verlauf gekennzeichnet. Insgesamt können die typischen Lebenszyklen von qualitativ minderwertigen und hochwertigen Artikeln mit dem bisherigen Ansatz nur teilweise erfasst werden.

2.3 Zeitreihenanalyse

Einen interessanten Ansatz um die Charakteristika der Lebenszyklen von qualitativ minderwertigen und hochwertigen Artikel ableiten zu können, bieten Verfahren zur Zeitreihenanalyse. Zeitreihen, wie der Lebenszyklus eines Artikels, sind jedoch vielfach nicht direkt verarbeitbar. So sind diese oft metrisch skaliert, was eine Verarbeitung für eine Vielzahl an Verfahren unmöglich bzw. schwierig macht. Um beispielsweise Algorithmen aus anderen Disziplinen wie der Textanalyse oder der Bioinformatik zu nutzen, werden Zeitreihen oft in eine symbolische Repräsentation überführt (Almeida et al. 2001, S. 429ff.). In der Vergangenheit wurde eine Vielzahl an Repräsentationen aller Art entwickelt. Einen kurzen Überblick hierzu bieten Lin et al. (2003) bzw. Daw et al. (2003).

Da eine Reduzierung der Dimension auch immer Verlust von Information bedeutet, sollte eine Repräsentation die Struktur bzw. Charakteristik der Daten nicht zerstören. So können Lebenszyklen einen sehr gleichmäßigen oder auch sehr chaotischen und wechselhaften Verlauf aufweisen. Diese „Charakteristik" wird durch bestehende Repräsentationen allerdings nur unzureichend erfasst. Darüber hinaus haben viele Verfahren den Nachteil, dass die transformierte und die originale Zeitreihe eine identische Länge besitzen (Lin et al. 2003). Dies erschwert die Vergleichbarkeit, die allerdings Grundlage für viele weitere Verfahren wie Klassifizierung oder Clustering ist. Auch Lebenszyklen weisen häufig unterschiedliche Längen auf, was die Anwendung bestehender Repräsentationen zur Unterscheidung von Lebenszyklen erschwert. Eine für das Anwendungsfeld der Lebenszyklusanalyse adäquate Repräsentation sollte daher Zeitreihen beliebiger Länge auf eine einheitliche Größe abbilden. Aufgrund der genannten Schwächen wird deshalb im vorliegenden Beitrag ein neues Verfahren zur Zeitreihenrepräsentation vorgestellt.

3 Modell zur Repräsentation von Zeitreihen auf Basis von Markov-Modellen

Das vorgestellte Verfahren zur Zeitreihenanalyse basiert auf zwei Schritten. Zunächst wird die Zeitreihe auf Basis der Anstiege der Messwerte in den einzelnen Zeitpunkten in eine symbolische Repräsentation überführt und in einem zweiten Schritt in ein Markov-Modell umgewandelt.

3.1 Symbolische Anstiegsbasierte Zeitreihenrepräsentation

Um die Charakteristik eines Lebenszyklus bestmöglich abzubilden, basiert das folgende Verfahren auf dem neuartigen Ansatz, die Änderung des Anstiegs in jedem Zeitpunkt der Zeitreihe in die Repräsentation mit einfließen zu lassen.

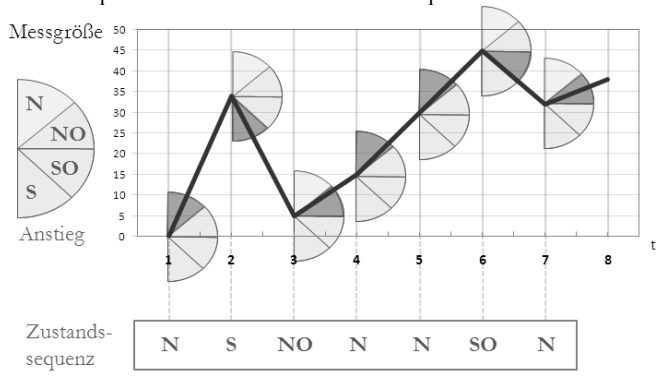


Abbildung 1: Anstiegsbasierte Repräsentation einer Zeitreihe[3]

[3] Zur besseren Übersicht wurden hier lediglich 4 mögliche Zustände verwendet. Eine reale Ausprägung kann eine unterschiedliche Zahl von Zuständen enthalten.

Wie in Abbildung 1 beispielhaft zu erkennen, werden zur Repräsentation die Anstiege zwischen zwei benachbarten Messzeitpunkten genutzt. Für jeden Zeitpunkt t wird der Anstieg *m* zweier benachbarter Punkte x_t und x_{t+1} wie folgt berechnet:

$$m = x_{t+1} - x_t$$

Danach wird der Wertebereich für alle *m* in *N* verschiedene disjunkte Abschnitte $S=(s_1,\ldots,s_N)$ unterteilt, was einer Kategorisierung bzw. Diskretisierung aller möglichen Anstiege in N Klassen entspricht. An jedem Messzeitpunkt befindet sich der Wert der Zeitreihe in genau einer dieser Anstiegsklassen aus *S*. Wie am Beispiel in Abbildung 1 zu sehen, können diese Klassen auch mithilfe verbaler Entsprechungen wie *Nord ("N"), NordOst ("NO"), SüdOst ("SO"), Süd ("S") etc.* interpretiert werden. Die gesamte Zeitreihe O kann für T Messpunkte als Folge von Anstiegsklassen abgebildet werden:

$O = X_0,X_1,\ldots,X_t,X_{t+1}\ldots,X_T$ Bsp.: $O =$ *"N","S","NO","N","N","SO","N"*

Wobei $X = (X_1,\ldots,X_T)$ eine Beobachtungssequenz von Zufallsvariablen ist, die jeweils einen Wert *S*, $|S| = N$ annehmen können. Eine so erzeugte Sequenz kann durch verschiedenste Algorithmen verarbeitet werden, beinhaltet aber dennoch die Charakteristik des Verlaufs der ursprünglichen Zeitreihe. Annahmegemäß ergeben sich für die Klassen *qualitativ hochwertig* und *minderwertig* charakteristische Verläufe. Dessen ungeachtet weisen Originalzeitreihe und symbolische Repräsentation immer noch eine identische Länge auf, was durch die Anwendung eines Markov-Modells im folgenden Abschnitt gelöst wird.

3.2 Anwendung des Markov-Modells auf Repräsentation

Markov-Modelle werden häufig dazu verwendet, eine Folge von Zustandswechseln in einem System zu modellieren. Abbildung 2 verdeutlicht ein vereinfachtes Modell bestehend aus zwei Zuständen „N" und „S". In jedem Zeitpunkt befindet sich das System in genau einem Zustand, aus diesem in einen Folgezustand gewechselt wird. Die Gewichte der ausgehenden Kanten beschreiben die entsprechende Wechselwahrscheinlichkeit. Wird beispielsweise der wiederholte Münzwurf mit zwei Zuständen abgebildet, so ergeben sich für eine faire Münze jeweils Kanten mit einer Gewichtung von genau 50%.

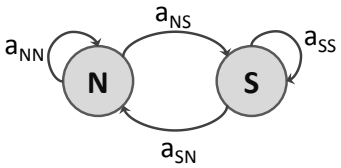


Abbildung 2: Beispiel eines Markov-Modells mit zwei Zuständen "N" und "S".

Zur Überführung einer symbolischen Zeichenkette in ein Markov-Modell wird der jeweilige Anstieg in Zeitpunkt *t*, also die jeweilige Klasse *s*, als Zustand interpretiert. Die Zeitreihe kann so als Folge von Zustandswechseln - als Markov-Prozess - abgebildet werden. Die Zeitreihe aus Abbildung 1 bietet beispielsweise maximal $N^2 = 16$ verschiedene Zustandswechsel, was 16 Kanten in Abbildung 2 entsprechen würde. So kann aus dem Zustand "N" in drei andere Zustände "NO","SO","S" gewechselt werden, oder im aktuellen Zustand „N" verharrt werden. Eine Summierung aller Zustandswechsel über alle Zeitpunkte s_t zu s_{t+1} kann in einer Matrix $A \in R^{N \times N}$ zusammengefasst werden. Diese Matrix wird als Übergangsmatrix bezeichnet. Die Übergangsmatrix hat die Form:

$$a_{ij} = P(X_{t+1} = s_j \mid X_t = s_i) \qquad \forall\, i,j \in S$$

Die Wahrscheinlichkeit des Übergangs einer Anstiegsklasse *i* in die Anstiegsklasse *j* innerhalb der Zeitreihe wird durch a_{ij} verdeutlicht. Die Übergangsmatrix spiegelt wider, mit welcher Wahrscheinlichkeit unterschiedliche Anstiegsklassen aufeinander folgen. Aus der Sequenz aus Abbildung 1 folgt beispielsweise, dass die Wahrscheinlichkeit von Zustand *N* in Zustand *S* zu wechseln 66% beträgt, da dieser Zustandswechsel zweimal für insgesamt drei Zustandswechsel ausgehend von N (Zeitpunkt 1, 4 und 5) beobachtet wurde.

Die so ermittelte Übergangsmatrix einer Zeitreihe spiegelt deren individuelle Charakteristik wider und kann als eine Art „Fingerabdruck" verstanden werden, welcher unabhängig von der Länge einer Zeitreihe über eine identische Anzahl an Parametern verfügt. Ein weiterer Vorteil ist, dass diese Matrix nicht nur für eine einzelne Zeitreihe – einen einzelnen Lebenszyklus – gebildet werden kann, sondern auch für eine Vielzahl an Zeitreihen. Es ergibt sich daraus eine einzelne Matrix, und folglich eine einzelne Charakteristik, für eine Gruppe von Zeitreihen. Dazu müssen lediglich die Zustandswechsel über alle Zeitreihen erfasst werden. Das abschließende Markov-Modell $\lambda = \lambda(A,S,\pi)$ wird durch weitere Parameter und Bedingungen vervollständigt. So existiert ein Startvektor π für die Ermittlung des initialen Zustands des Modells.[4]

4 Automatische Qualitätsmessung mittels Zeitreihenanalyse

4.1 Daten und Methodik

Um die Effektivität von Messgrößen zur Qualitätsmessung zu bewerten, wird in diesem Beitrag die Güte der Messgrößen anhand des tatsächlichen Datenbestandes der Wikipedia evaluiert. Hierzu wurde der Datenbestand vom 21.01.2008 herunter

[4] Eine detaillierte Übersicht bietet Manning und Schütze (2005, S. 317-339).

geladen, der neben den Artikelversionen auch die gesamte Änderungshistorie enthält.[5]

Die Evaluationsmethodik entspricht dem in der Literatur verbreiteten Ansatz. Es werden zunächst anhand der Wikipedia-internen Bewertungen qualitativ minderwertige (Löschkandidaten) und hochwertige (Lesenswerte und Exzellente) Artikel aus dem Datenbestand selektiert. Im Rahmen der Wikipedia-internen Bewertung werden die Artikel auf zentralen Seiten gelistet (Liste der Löschkandidaten, Liste der Lesenswerten Artikel, Liste der Exzellenten Artikel). Dadurch steigt die Wahrnehmung der Artikel in der Community, sodass die Bearbeitungsintensität (insbesondere bei Löschkandidaten) zumeist stark zunimmt. Um die Evaluation von diesem Effekt zu bereinigen, werden die Lebenszyklen im Monat vor der Kandidatur zur jeweiligen Wikipedia-Bewertung abgeschnitten. Es wird davon ausgegangen, dass sich die Qualität im Vergleich zum Bewertungszeitpunkt kaum verändert. Einige Lebenszyklen werden dadurch stark verkürzt. Da eine Lebenszyklusanalyse nur ab einer Mindestlänge des Lebenszyklus sinnvoll ist, wurden alle Artikel mit einer Lebensspanne von weniger als 10 Monaten aus dem Datenbestand eliminiert.

Nach der Bereinigung des Datenbestands standen 113 minderwertige Artikel und 2354 hochwertige Artikel zur Verfügung. Um beide Gruppen zu gleichen Anteilen zu berücksichtigen, wurden aus der Gruppe der qualitativ hochwertigen Artikel 113 Artikel zufällig zur Analyse ausgewählt. Dieser Vorgang wurde für insgesamt drei Stichproben wiederholt. Durch den Vergleich der Evaluationsergebnisse aus den unterschiedlichen Stichproben kann die Repräsentativität der Untersuchung validiert werden. Da sich insbesondere die persistenten Änderungen zur automatischen Qualitätsmessung eignen (Wöhner und Peters 2009), beschränkt sich dieser Beitrag auf diese Änderungen. Für die Stichproben wurden aus der Änderungshistorie die persistenten Änderungen wie in Abschnitt 2.2 beschrieben berechnet. Eine Beurteilung der Güte der Zeitreihenanalyse erfolgt, indem auf deren Basis eine Klassifikation zwischen qualitativ minderwertigen und hochwertigen Artikeln durchgeführt wird. Anhand der Trefferquote lässt sich die Effektivität abschätzen.

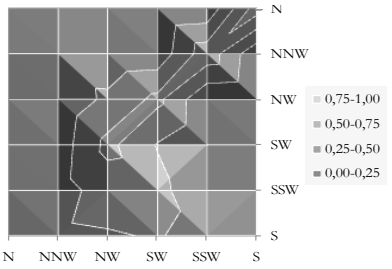
4.2 Klassifikation anhand der Lebenszyklusanalyse

Bevor die Klassifikation des Lebenszyklus eines Artikels vorgenommen werden kann, bedarf es der Ermittlung der typischen Lebenszyklen für die beiden Gruppen *qualitativ minderwertig* und *hochwertig*. Anhand dieser Referenzen kann dann ein zu testender Lebenszyklus klassifiziert werden. Hierzu wird zunächst jede Stichprobe in ein Trainingsset (2/3 der Artikel) und ein Testset (1/3 der Artikel) unterteilt. Für das Trainingsset wird, wie in Abschnitt 3.2 beschrieben, pro Stichprobe ein Markov-Modell λ_h, für hochwertige Artikel, und λ_m für minderwertige Artikel

[5] http://download.wikimedia.org/backup-index.html

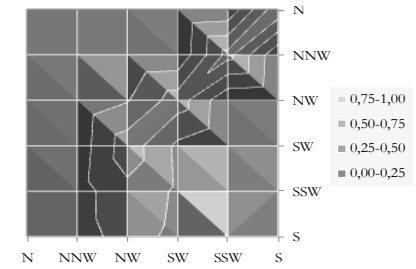
erstellt. Es werden hierbei sechs Anstiegsklassen berücksichtigt, da so die besten Trefferquoten erzielt werden konnten.

Übergangsmatrix "qualitativ hochwertig"



	S	SSW	SW	NW	NNW	N
S	0,02	0,16	0,27	0,23	0,21	0,12
SSW	0,02	0,05	0,33	0,37	0,14	0,09
SW	0,00	0,01	0,23	0,57	0,11	0,07
NW	0,00	0,03	0,56	0,30	0,06	0,04
NNW	0,06	0,65	0,08	0,06	0,08	0,07
N	0,87	0,05	0,01	0,01	0,03	0,02
Σ	0,97	0,95	1,49	1,53	0,64	0,42

Übergangsmatrix "qualitativ minderwertig"



	S	SSW	SW	NW	NNW	N
S	0,01	0,25	0,18	0,36	0,13	0,06
SSW	0,09	0,08	0,21	0,44	0,14	0,05
SW	0,01	0,02	0,20	0,64	0,10	0,04
NW	0,01	0,04	0,63	0,23	0,07	0,03
NNW	0,05	0,84	0,04	0,02	0,03	0,02
N	0,93	0,05	0,00	0,01	0,01	0,00
Σ	1,09	1,28	1,26	1,70	0,47	0,19

Abbildung 3: Übergangsmatrizen für Lebenszyklen der Gruppen "hochwertig" und "minderwertig"

Abbildung 3 veranschaulicht die erstellten Übergangsmatrizen für die beiden Gruppen *qualitativ hochwertig* und *minderwertig*. Die Lebenszyklen beider Gruppen sind gekennzeichnet durch stark schwankende Verläufe, was hohe Werte der Zustandsübergänge im Bereich „N→S" deutlich machen. Dennoch fällt auf, dass Lebenszyklen der Gruppe *qualitativ hochwertig* tendenziell stärker in Zustände einer steigenden Bearbeitung übergehen und somit häufiger intensiv bearbeitet werden.
Anstelle einer Verwendung der kompletten Matrix als Messgröße, bietet sich die Möglichkeit die Funktionalität des Markov-Modells zu nutzen, indem für jede mögliche Sequenz O die Wahrscheinlichkeit berechnet werden kann, dass O auf dem Markov-Modell $\lambda = \lambda(A,S,\pi)$ beobachtet wurde:

$$P(O \mid \lambda) = \pi_{X_1} \prod_{t=1}^{T-1} a_{X_t X_{t+1}}$$

Als Sequenz O wird der wie in Abschnitt 3.1 in eine Zeichenkette transformierte Lebenszyklus verstanden. So gibt beispielsweise $P(O \mid \lambda_m)$ die Wahrscheinlichkeit wider, dass die Beobachtung des Lebenszyklus O auf dem Modell der qualitativ minderwertigen Artikel auftritt. Ergebnis ist eine äußerst kompakte Messgröße für einen Artikel, welche die Charakteristik des Verlaufs operationalisiert. Zur Klassifikation des Lebenszyklus eines Artikels müssen lediglich die Wahrscheinlichkeiten der Beobachtung dieses Lebenszyklus $P(O \mid \lambda_h)$ und $P(O \mid \lambda_m)$ auf den beiden Model-

len λ_h und λ_m miteinander verglichen werden. Je größer die Wahrscheinlichkeit für ein Markov-Modell, desto ähnlicher ist der Verlauf des Lebenszyklus dem Verlauf der entsprechenden Gruppe.

Tabelle 1: Evaluationsergebnisse

Metrik	Gesamtgüte	Güte in Klasse hochwertig (TP)	Güte in Klasse minderwertig (TN)	Korrelation zur Diskriminanzfunktion
Markov-Modell Stichprobe 1	**86%**	87%	85%	Ø 0,735
Markov-Modell Stichprobe 2	**90%**	92%	87%	
Markov-Modell Stichprobe 3	**87%**	90%	84%	
Ergebnisse der Messgrößen aus Wöhner und Peters (2009)				
Mittelwert Persistente Änderungen M^{Per}	**87%**	76%	97%	Ø 0,414
Durchschnittliche Persistente Änderungen A^{Per}	**86%**	78%	94%	Ø 0,389
Länge des Artikels L	**82%**	77%	87%	Ø 0,436

Tabelle 1 zeigt die erzielten Klassifikationsergebnisse der Testsets sowohl für eine gesamte Stichprobe als auch in den einzelnen Gruppen. Es wird deutlich, dass die auf Markov-Modellen basierende Messgröße geeignet ist die Qualität eines Artikels zu klassifizieren. Mit einem Wert zwischen 86%-90% erreicht die Messgröße sehr hohe Werte, vergleichbar beispielsweise mit den Ergebnissen die Wöhner und Peters (2009) in ihrer Untersuchung für andere Messgrößen ermitteln konnten. Interessant ist außerdem, dass die hier vorgestellte Messgröße in beiden Gruppen eine ähnlich hohe Klassifikationsgüte aufweist (True Positive TP, True Negative TN) wogegen bisherige Messgrößen in den Gruppen unterschiedlich gute Vorhersagen zeigten.

4.3 Kombination mit etablierten Messgrößen

Die Ergebnisse aus Tabelle 1 werfen die Frage auf, inwieweit eine Kombination mehrerer Messgrößen die Klassifikationsgüte zusätzlich steigern kann. Wöhner und Peters (2009) haben eine Vielzahl an möglichen Messgrößen untersucht, bisher erfolgte jedoch keine Kombination dieser Messgrößen. Aus diesem Grund wurden diese Messgrößen für die vorliegenden Daten berechnet und in Kombination mit der hier vorgestellten Messgröße in einer Diskriminanzanalyse untersucht. Die von Fisher (1936) entwickelte Diskriminanzanalyse als multivariates Verfahren ermöglicht eine Unterscheidung von Gruppenzugehörigkeiten und kann so zur Klassifikation von Elementen verwendet werden. Innerhalb der Diskriminanzanalyse wird eine Diskriminanzfunktion formuliert, die jeder Merkmalsvariable (hier Messgröße) einen Koeffizienten bzw. ein Gewicht zuordnet.

Darüber hinaus kann über die Korrelation der Messgrößen mit der Diskriminanzfunktion der Beitrag jeder einzelnen Variable zur Erklärung der Gruppenunterschiede ermittelt werden.

Durch die Kombination der Messgrößen konnte die Trefferquote der Klassifikation auf durchschnittlich 90% gesteigert werden. Auch zeigen die jeweiligen Korrelationen der Messgrößen zur Diskriminanzfunktion in Tabelle 1, dass die auf Markov-Modellen basierende Messgröße den größten Anteil zur Erklärung der Gruppenzugehörigkeit liefert. Mit einem Wert von 0,735 liegt diese deutlich über den Werten anderer Messgrößen. Dennoch ist anzumerken, dass eine Kombination mehrerer Messgrößen das Klassifikationsergebnis nur geringfügig verbessert.

5 Schlussbetrachtung und Ausblick

Dieser Beitrag untersuchte die automatische Qualitätsmessung von Wikipedia-Artikeln anhand einer neuartigen auf einer Lebenszyklusanalyse basierenden Messgröße. Wie gezeigt werden konnte, bietet das vorgestellte Verfahren zur Zeitreihenanalyse die Möglichkeit einer kompakten und vielseitigen Repräsentation von Lebenszyklen, und dies ohne einen Informationsverlust an charakteristischen Eigenschaften. Das Verfahren ist überdies in der Lage, Zeitreihen jeglicher Art auf eine einheitliche Messgröße zu operationalisieren, die dann durch Standardverfahren verarbeitet werden kann. Eine Evaluierung der Messgröße anhand einer Klassifikation zeigte, dass durch die vorgestellte Zeitreihenanalyse bestehende Verfahren zur Qualitätsmessung in der Wikipedia verbessert werden konnte.

Eine potentielle Schwäche der bestehenden Forschung ist das Kriterium der Wikipedia-internen Bewertung der Artikel durch die Community, insbesondere da der Begriff der „Qualität“ oft von subjektiver Natur ist. Alternative Methoden wie Expertenratings können hier in zukünftigen Untersuchungen als Referenz dienen. Forschungsbedarf besteht außerdem in der Frage der Robustheit unterschiedlicher Messgrößen, insbesondere von Messgrößenkombinationen, hinsichtlich einer Manipulation. Bedingt durch die offene Bearbeitung in Wikis ist dies ein wichtiger Aspekt der automatischen Qualitätsmessung. Innerhalb der Änderungshistorie von Wikipedia-Artikel ist eine hohe Menge an Informationen enthalten, die es in Zukunft zu untersuchen gilt. Auf Basis des vorgestellten Verfahrens sind weitere Untersuchungen beispielsweise zur Mustererkennung denkbar. Zukünftige Arbeiten könnten die Erkennung typischer bzw. auffälliger Lebenszyklen untersuchen. Festzuhalten bleibt, dass die Lebenszyklusanalyse einen lohnenden Ansatz zur Untersuchung einer Qualitätsmessung in der Wikipedia darstellt.

Literatur

Adler BT, de Alfaro L (2007) A content-driven reputation system for the Wikipedia. In: Proceedings of the 16th International Conference on the World Wide Web. Banff.

Adler BT., Chatterjee K, de Alfaro L, Faella M, Pye I.and Raman V (2008) Assigning trust to Wikipedia content. In: Proceedings of the 2008 International Symposium on Wikis. Porto.

Almeida JS, Carriço JA, Maretzek A, Noble PA, Fletcher M (2001), Analysis of genomic sequences by Chaos Game Representation. Bioinformatics 17(5): 429-437

Blumenstock JE (2008) Size matters: word count as a measure of quality on Wikipedia. In: Proceedings of the 17th international conference on World Wide Web. Beijing.

Cross T (2006). Puppy smoothies: Improving the reliability of open, collaborative wikis. In: First Monday, 11(9).

Cunningham W, Leuf B (2001) The Wiki Way. Quick collaboration on the web. Addison-Wesley, Boston.

Daw CS, Finney DEA, Tracy ER (2003) A review of symbolic analysis of experimental data. Review of Scientific Instruments, Vol. 74 (2): 915-930.

Denning P, Horning J, Parnas D, Weinstein L (2005) Wikipedia Risks. In: Communications of the ACM. 48(12):152.

Dondio P, Barrett S (2007) Computational trust in web content quality: a comparative evalutation on the Wikipedia project. In: Informatica – An International Journal of Computing and Informatics 31(2):151-160.

Fisher RA (1936) The use of multiple measurements in taxonomic problems. Annals Eugenics 7:179-188

Hippner H, Wilde T (2005) Social Software. In: WIRTSCHAFTSINFORMATIK 47(6): 441-444.

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.

Lim EP, Vuong BQ, Lauw HW, Sun A (2006) Measuring qualities of articles contributed by onlinecommunities. In: Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence. Hong Kong.

Lin J, Keogh E, Lonardi S, Chiu B (2003) A symbolic representation of time series, with implications for streaming algorithms. In: Proceedings of the 8th ACM SIGMOD workshop on Research issues in data mining and knowledge discovery (2003), 2-11, San Diego

Manning CD, Schütze H (2005) Foundations of statistical natural language processing, 8. Ausgabe. MIT Press, Cambridge

O'Reilly T (2005). What is Web2.0? - Design Patterns and Business Models for the Next Generation of Software. http://oreilly.com/web2/archive/what-is-web-20.html. Abruf am 2009-09-02.

Stvilia B, Twidale MB, Smith LC, Gasser L (2005) Assessing information quality of a community-based encyclopedia. In: Proceedings of the International Conference on Information Quality. Cambridge.

Víegas F, Wattenberg M, Dave K (2004) Studying cooperation and conflict between authors with history flow visualizations. In: Proceedings of the SIGCHI Conf. on Human Factors in Computing Systems. Vienna.

Wikipedia. (2009). Autorenportal. http://de.wikipedia.org/wiki/Wikipedia:Autorenportal. Abruf am 2009-09-02.

Wöhner T, 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.

Zachte E (2009) Wikipedia Statistics. http://stats.wikimedia.org/EN/Sitemap.htm. Abruf am 2009-09-04.

Zeng H, Alhoussaini M, Ding L, Fikes R., McGuinness D (2006) Computing trust from revision history. In: Proceedings of the Intl. Conf. on Privacy, Security and Trust. Markham.