Philipp Scholl, Doreen Böhnstedt, Renato Dominguez Garcia, Christoph Rensing, Ralf Steinmetz: *Anwendungen und Nutzen der Automatischen Erkennung von Web-Genres in persönlichen und Community-Wissensnetzen*. In: Andreas Schwill, Nicolas Apostopoulos: Lernen im digitalen Zeitalter - Workshop-Band - Dokumentation der Pre-Conference zur DeLFI 2009, p. 37-44, Logos, September 2009. ISBN 9783832522735.

Anwendungen und Nutzen der Automatischen Erkennung von Web-Genres in persönlichen und Community-Wissensnetzen

Philipp Scholl, Doreen Böhnstedt, Renato Dominguez Garcia, Christoph Rensing, Ralf Steinmetz

Multimedia Kommunikation (KOM), Technische Universität Darmstadt
64283 Darmstadt
http://www.kom.tu-darmstadt.de
{scholl, boehnstedt, renato, rensing, ralf.steinmetz}@kom.tu-darmstadt.de

Zusammenfassung: Inhalte aus dem Web werden für das Ressourcen-basierte Lernen immer wichtiger. ELWMS.KOM bietet Lernenden die Möglichkeit, Webressourcen in ihren persönlichen Wissensnetzen zu persistieren und organisieren. Mehrere persönliche Wissensnetze können zu einem großen Community-Wissensnetz aggregiert werden. Dabei ist wichtig, dass Lernende ihre Webressourcen wieder auffinden können. Neben thematischen Aspekten bietet die Art der Informationen in einer Webressource den Lernenden dafür gute Anhaltspunkte. Die Informationsart kann auf das Web-Genre, also den Typ einer Webressource, abgebildet werden. Wir zeigen in dieser Arbeit, dass eine automatische Web-Genre-Erkennung sinnvoll und durchführbar ist, präsentieren eine Evaluation und stellen die Integration in ELWMS.KOM dar.

1 Motivation und Einführung

Das Internet entwickelt sich immer mehr zu einem überall zugänglichen und alltäglichen Medium, das zunehmend unsere Arbeits- und Lebensweisen durchdringt. Es finden sich viele Ressourcen, die relevante Informationen enthalten und zu Lernzwecken dienen können. Ein Beispiel hierfür sind Webressourcen aus Wikis, Foren und Blogs, die oft zur Erstellung von Inhalten eingesetzt werden und dementsprechend möglicherweise relevante Informationen beinhalten. Oft sind diese Ressourcen nicht explizit für Lern- oder Lehrzwecke intendiert, können jedoch von Lernenden dazu benutzt werden. Dieses *Ressourcen-basierte Lernen* findet nicht zwangsweise in Bildungsinstitutionen statt, sondern geschieht oft aus eigenen Interessen heraus und ist somit frei und selbstgesteuert. Mit dem damit verbundenen Wegfall von Autoritäten wie Lehrern oder Tutoren ist der Lernende jedoch selbst der Organisator seines Lernprozesses und muss selbständig viele verschiedene Prozesse durchführen: der anfängliche Informationsbedarf muss identifiziert werden und relevante Ressourcen müssen bewertet und ausgewählt werden. Weiterhin muss der Lernende diese in seine vorhandene Wissensstruktur einordnen und – wenn die gefundenen Informationen festgehalten werden sollen – wiederauffindbar speichern. Oft ist eine Wiederverwendung der Informationen wünschenswert, um sie anderen Interessierten weitergeben zu können.
ELWMS.KOM ist ein Wissensnetz für die Organisation von Webressourcen für das Ressourcen-basierte Lernen, das diese Prozesse adressiert (siehe [BS08a]). Dabei steht der Aspekt des persönlichen Wissensnetzes im Vordergrund, um Lernenden die Möglichkeit zu geben, gefundene, für sie relevante Webressourcen in ihre persönliche Wissensstruktur einzuordnen. Weiterhin können sie ihre Ressourcen typisiert taggen, d.h. den Ressourcen Schlüsselworte zuordnen, die gleichzeitig mit einem Typ belegt werden können (z.B. als Thema, Ereignis, Ort oder Person, siehe Abb. 1) [BS08b]. Damit können Kontextinformationen einer Ressource abgebildet werden und ermöglichen später über diese den Zugriff. Inhalte können nach Tagtypen

gefiltert werden, d.h. es werden beispielsweise nur Personen angezeigt.

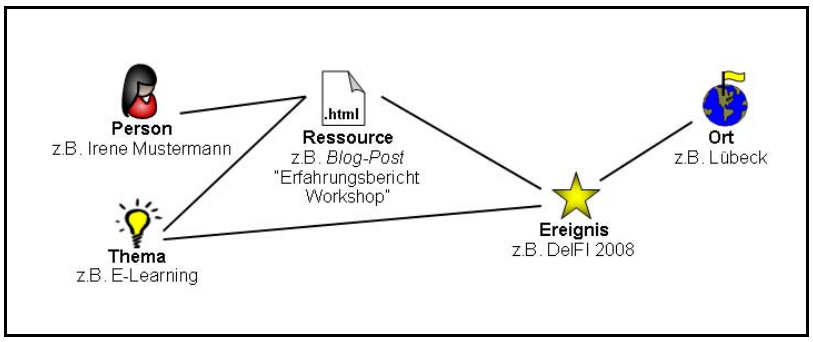

Abb. 1 Ausschnitt aus einem persönlichen Wissensnetz, das eine Webressource mit verschiedenen Tag-Typen verknüpft

Wissensnetze können mit der Zeit sehr umfangreich werden, was das Wiederfinden bereits gespeicherter Webressourcen beeinträchtigen kann. Deshalb ist es notwendig, geeignete Retrievalmöglichkeiten bereitzustellen, die es Lernenden ermöglichen, nach ihren Lernressourcen effektiv und effizient zu suchen. Dieses Problem verschärft sich sogar noch, wenn viele persönliche Wissensnetze in ein Community-Wissensnetz aggregiert werden. Hier ist eine geeignete Unterstützung des Retrievals unumgänglich, da das Community-Wissensnetz als umfangreiche Quelle für bereits von anderen Lernenden als relevant erachtete Informationsartefakte dienen kann.
ELWMS.KOM unterstützt textbasiertes Retrieval (also eine Suche nach Stichwörtern), aber es gibt auch andere Eigenschaften von Webressourcen, die für Suchende nützlich sind. Hat z.B. der Lernende seinen Informationsbedarf identifiziert und sich Ziele für seine Recherche gesteckt, ist ihm zumeist auch klar, was für eine Art von Information (z.B. Überblickswissen, Diskussionen, Faktensammlungen zu einem bestimmten Thema) er benötigt, um diesen zu decken [Va00]. Somit kann für Lernende beim Retrieval das Wissen über die Informationsart einer Ressource wichtig sein. Oft lassen sich diese Informationsarten auf bestehende Software-Systeme im Web zurückführen. Systeme wie Blogs, Wikis oder Foren entstanden aus dem Bedürfnis heraus, spezielle Informationsarten abzubilden. Sie sind Beispiele für unterschiedliche *Web-Genres*, also typische Ausprägungen von Webressourcen, die sich durch Struktur, Funktionalität, Art der Nutzung und Inhalte unterscheiden. So sind z.B. typische Inhalte für Foren Diskussionen, die oft mit einer initialen Frage bzw. einem Problem angestoßen wurden. In Wikis werden häufig Faktensammlungen oder Vorgehen wie Tutorials festgehalten (die oft in kollaborativer Arbeit entstehen) und Blogs dienen oft dem persönlichen Festhalten und Darstellen von Meinungen, Gelerntem oder aktuellen persönlichen „Projekten“.
Somit ist es für ein System wie ELWMS.KOM, das ein Persistieren und ein Retrieval von Artefakten solcher Web-Genres unterstützt, nützlich, das Wissen über das Web Genre einer Ressource als weitere Kontextinformation zu speichern und in die Suche einzubeziehen, so ist z.B. Suchenden oft bekannt, dass die gesuchte Information sich etwa in einem Blogpost befand.
In dieser Arbeit werden wir wie folgt vorgehen: In Kapitel 2 wird näher auf verwandte Arbeiten zur Bedeutung von Webressourcen für das Lernen und Genre-Informationen im Retrieval eingegangen. Kapitel 3 fokussiert Ansätze der Web-Genre-Erkennung, Kapitel 4 führt in die grundlegende Funktionsweise unseres Ansatzes ein und stellt unseren Evaluationscorpus vor. In Kapitel 5 präsentieren wir die Ergebnisse einer Evaluation gefolgt von der Implementierung in Kapitel 6 und schließen mit einer Zusammenfassung und einem Ausblick in Kapitel 7.

2 Bedeutung von Webressourcen fürs Lernen und Nutzen von Genre Informationen im Retrieval

Für Retrieval im eigenen Wissensnetz sind Kontextinformationen sehr relevant für das Wiederauffinden von bestimmten Informationen. Dabei spielt das Wissen über im Kontext getätigte Aktivitäten eine große Rolle [NP05], als auch näheres Wissen um Eigenschaften der Ressource selbst. Insbesondere das Genre einer Ressource hilft, den Nutzen der in dieser Ressource enthaltenen Informationen besser einzuordnen [YS97].
Eine Studie von Vakkari [Va00] bestätigt, dass das Bedürfnis von Lernenden nach bestimmten Informationsarten von verschiedenen Voraussetzungen bestimmt wird. So suchen Lernende zwar in erster Linie nach Dokumenten, die thematisch ihrem Informationsbedarf genügen, in zweiter Präferenz jedoch Ressourcen, die einer bestimmten Informationsart entsprechen, von der die Lernenden sich einen hohen Nutzen für die Lösung ihres Problems erwarten. Weiterhin ändern sich mit dem Voranschreiten der Recherche auch die Anforderungen an die Informationsart einer Ressource; so benötigen z.B. Lernende in den Endphasen einer Recherche weniger Hintergrundinformationen und ziehen fein granulare, klar fokussierte Informationen vor.
In [Va00] werden ausschließlich wissenschaftliche Ressourcen in Betracht gezogen, die in einer existierenden Datenbank bereitgestellt werden. Beim informellen Lernen gewinnen jedoch Webressourcen als Informationsquellen an Bedeutung, da sie von überall her zu jeder Zeit zugänglich sind.
Speziell für das Retrieval von Webressourcen zeigt eine Studie von [MS04], dass die Suche nach Informationen auch stark vom Bedürfnis nach bestimmten Informationsarten bestimmt wird. Suchende haben gewisse Erwartungen an verschiedene Genres von Webressourcen, die zum Teil von deren funktionalen Eigenschaften abhängen. Dabei zeigt die Studie, dass Suchende bestimmte Web-Genres mit Erwartungen an die Nützlichkeit und Relevanz einer Webressource verbinden. So sagen 93% der 284 Teilnehmer aus, dass sie eine Filterung ihrer Suchergebnisse nach Web-Genres für „sehr nützlich“ oder „oft nützlich“ halten.
Da Benutzer üblicherweise Metadaten, für die sie keinen unmittelbaren Mehrwert sehen, nicht explizit spezifizieren, ist eine automatische Erkennung des Web-Genres (Web Genre Detection) in Systemen wie ELWMS.KOM für das Retrieval wünschenswert.

3 Automatische Web-Genre-Erkennung

Seit den späten 80ern Jahren war die automatische Klassifizierung von elektronischen Texten ein großes Forschungsgebiet. Da diese frühen Ansätze sich meistens nur auf die Klassifizierung von Texten und Textarten fokussierten, beschränkten sie sich auf die linguistische Analyse und ausgewählte strukturelle Metriken (Satzzeichenfrequenzen, Satzlängen und Lesbarkeitsmaße wie die Flesch-Metrik). Mit dem Aufkommen des Webs wurde die Erkennung von Web-Genres immer wichtiger, die vor allem die funktionalen Eigenschaften von Ressourcen im Web und deren Struktur in den Mittelpunkt stellt.
Methoden zur automatischen Web-Genre-Erkennung werden vor allem in Retrieval-Szenarien eingesetzt und sind kein gänzlich neues Forschungsfeld, wie einige bereits existierende Suchmaschinen zeigen, die sich auf konkrete Web-Genres spezialisieren, wie z.B. Google Blogs für Weblogs oder Google Scholar für wissenschaftliche Ressourcen.
Es gibt dafür verschiedene Ansätze, die – je nachdem, welche Ziele sie verfolgen – unterschiedliche Web-Genres als relevant definieren, auf unterschiedliche Art und Weise Features aus den Webressourcen gewinnen und verschiedene Ansätze der Klassifikation verfolgen.
Meyer zu Eissen et al. [MS04] identifizieren – basierend auf einer Nutzerstudie – relevante Web Genres wie Hilfeseiten (z.B. FAQs), Artikel (hauptsächlich wissenschaftliche Artikel sowie Seiten, die längere Texte enthalten), Diskussionen (Foren und Mailinglisten), Online-Shops, Linkseiten, Downloadseiten für Software sowie private und nicht-private (z.B. von Firmen oder Organisationen) Homepages. Da für Meyer zu Eissen et al. die in diesen Genres verwendete

Sprache ein wichtiges Unterscheidungsmerkmal ist, werden neben rein strukturellen Features (wie z.B. Häufigkeiten bestimmter funktionaler Elemente von HTML-Seiten wie Formulare, Tabellen und Überschriften), einfachen textstatistischen Features (wie z.B. Anzahl der Sätze oder Satzzeichenhäufigkeiten) und Worthäufigkeitsklassen vor allem linguistische Features wie Part-Of-Speech Analyse und syntaktische Analyse eingesetzt. Weiterhin wird die Häufigkeit von genre-typischen Worten (closed-class word sets) erhoben. Auf ihrem Corpus erreichen sie mit Support Vector Machines (SVM) eine Erkennungsgenauigkeit von durchschnittlich 70%.
Santini [Sa07] identifiziert die Genres Blogs, e-Shops, FAQs, Online-Frontpage (Hauptseiten von Institutionen), Linklisten, persönliche Homepages und Suchmaschinenseiten. Sie erweitert die Features von [MS04] um Häufigkeiten von HTML-Elementen und funktionale Klassen von HTML, die nach einer Evaluation besonders ausgeprägt für ihre respektiven Web-Genres sind. Die Erkennungsgenauigkeit dieses Ansatzes für Single-Label-Klassifizierung liegt bei 88-90% bei einem genau auf diese Genres abgestimmten Corpus.
Elgersma et al. [ER06] beschränken sich auf die Erkennung von Weblogs und unterscheiden damit ein spezielles Genre von Outliern (also andere Webseiten, die keinem der betrachteten Genres entsprechen). Dafür werden Features entwickelt, die die Struktur (z.B. basierend auf der Anzahl von Kommentaren, Vorhandensein von CSS-Dateien und RSS-Feeds), den Inhalt (z.B. closed-class word-sets) oder die Herkunft (Liste von bekannten Blog-Hosting-Services) repräsentieren. Mit einer Auswahl dieser Features erreichen sie bis zu 93% Genauigkeit bei der Klassifizierung mit verschiedenen Machine Learning Algorithmen.

4 Umsetzung der Web-Genre-Erkennung

Der Schwerpunkt dieser Arbeit ist es, zwischen den drei Web Genres Blog, Wiki und Forum zu unterscheiden. Zusätzlich betrachten wir Outliers, was in anderen Ansätzen oft vernachlässigt wird. Da sich die Startseiten in Blogs und Foren oft strukturell stark von den Seiten, auf denen der spezifische Inhalt dargestellt wird, unterscheiden, sind diese beiden Genres jeweils in zwei Sub-Genres aufgeteilt.

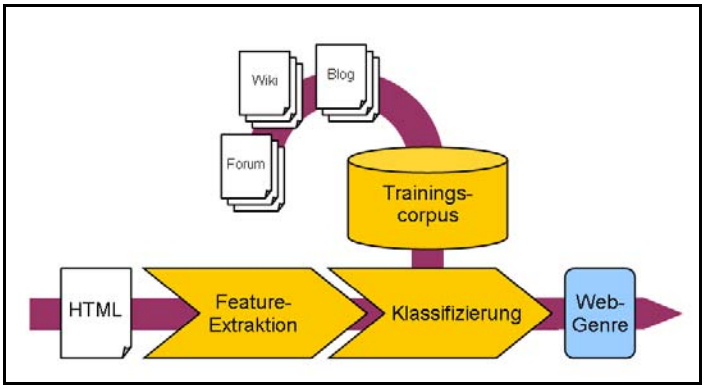

Abb. 2 Schematischer Ablauf der Web-Genre-Erkennung

Eine Besonderheit unseres Ansatzes ist der Verzicht auf eine linguistische Analyse der Texte auf den Seiten, da – zumindest bei den betrachteten Genres – allein die Struktur und der Aufbau einer Webressource bereits Aufschluss auf das zugrundeliegende Genre geben können. Dies bedeutet, dass die Erkennung der hier genannten Web Genres im Gegensatz zu verwandten Ansätzen unabhängig von der verwendeten Sprache durchgeführt werden kann.
Die Genre-Erkennung durchläuft die typischen Schritte Feature-Extraktion und Klassifizierung (siehe Abb. 2). Bei der Feature-Extraktion wird das vorhandene HTML-Markup analysiert und die Features werden aufgrund der erhobenen Eigenschaften berechnet. Eine Support Vector Machine, die mit manuell klassifizierten Beispieldaten trainiert wurde, bestimmt auf deren Basis das wahrscheinlichste Genre der Webressource. Im Folgenden werden die verwendeten Features näher erklärt und der Corpus der Trainingsdaten vorgestellt.

4.1 Features

Die in unserem Ansatz erhobenen Features (insgesamt 338) sind teils den verwandten Arbeiten entlehnt, teils selbst hergeleitet. Eine genauere Beschreibung der Herleitung und Funktionsweisen findet sich in [DS08]. Die Features lassen sich grob in folgende Kategorien einordnen:

Features, die aufgrund einfacher Markup-Statistiken ermittelt werden

Wie [Sa07] zeigt, sind Features, die die Häufigkeiten des Vorkommens von bestimmten HTML-Elementen repräsentieren ein gutes Maß für das Erkennen eines Genres. Dies lässt sich dadurch erklären, dass verschiedene Genres unterschiedlichen Zwecken dienen und dadurch auch unterschiedliche Darstellungen bzw. Interaktionsmöglichkeiten mit dem Inhalt ermöglichen. So ist es z.B. für eine Blogseite typisch, dass Kommentare hinterlassen werden können. Somit ist das Vorkommen von HTML-Elementen, die Eingabemöglichkeiten repräsentieren, hier verstärkt zu erwarten. Weiterhin werden Features erhoben, die das Verhältnis von HTML-Markup und Text beschreiben und die die Größe einer Webressource in Worten repräsentieren.

Features, die auf Verlinkung und URL-Eigenschaften basieren

Wichtige Kriterien für die Unterscheidung zwischen Outliers und unseren fokussierten Genres sind die Anzahl von verlinkten RSS-Feeds, die Struktur der URL für die analysierte Seite und die Art der aus der Seite herausführenden Links. So ist z.B. für viele längere Wiki-Seiten typisch, dass anfangs ein Inhaltsverzeichnis generiert wird, das auf alle folgenden Abschnitte mit seiten-internen Links verweist. Foren, bei denen sich die Mitglieder registrieren müssen, weisen oft Links auf dieselbe Webpräsenz auf, die auf die Profile der jeweiligen Autoren verweisen, wohingegen bei Blogs die Autoren-Links bei Kommentaren meistens extern auf die jeweiligen Blogs verweisen.

Features, die die Struktur der Webseite repräsentieren

Ein Großteil der Informationen im Web besteht aus von Web-Applikationen generierten Inhalten. Dies bedeutet, die Informationen werden von einem System aufbereitet und in HTML-Templates präsentiert. Bei Inhalten, die aus mehreren Artefakten bestehen und aus einer Datenbank ausgelesen werden (wie z.B. Kommentare zu einem Blogpost, siehe Abb. 3), wiederholen sich bestimmte Strukturen auf den verschiedenen Seiten, *Patterns* genannt.

Diese lassen sich durch eine Analyse der hierarchischen Struktur des Document Object Models (DOM, eine Baum-Repräsentation des HTML Markups) bestimmen, wobei Patterns derselben Ausprägung leichte Unterschiede in ihrer Struktur haben dürfen (z.B. in Kommentaren, da manche Autoren zusätzliches HTML-Markup wie Auszeichnung von hervorgehobenem Text oder Hyperlinks einfügen). Für eine ausführlichere Beschreibung des Extraktions-Algorithmus verweisen wir auf [DS08].

Abb. 3 Beispiel für ein Pattern: Kommentare zu einem Post in einem Weblog

Aus den Patterns lassen sich Features extrahieren, die die logische Struktur einer Webressource repräsentieren, wie z.B. die Anzahl von Patterns, ihre durchschnittliche Größe und das Verhältnis von Inhalten in Patterns zu „un-patterned“ Inhalten.

4.2 Der Trainingscorpus

Machine Learning Algorithmen benötigen einen Trainingscorpus, um an Beispielen zu lernen und die Features zu messen, die für verschiedene Genres charakteristisch und dementsprechend ausgeprägt sind. Für die automatische Web-Genre-Erkennung benötigten wir daher einen Corpus, der korrekt klassifizierte Beispiele aller für uns relevanten Genres enthält.
Für einen realistischen Corpus, der eine Untermenge des Webs repräsentieren soll, gelten darum einige Anforderungen. So soll der Corpus Stichproben von unterschiedlichen, aktuellen und realen Seiten enthalten. Diese dürfen in unterschiedlichen Sprachen verfasst sein und sollten von verschiedenen Anwendungen (z.B. im Fall von Blogs durch Wordpress, Blogger.com und anderen Blogging-Applikationen, die weniger verbreitet sind) generiert worden sein, um eine heterogene Auswahl von Seiten zu bekommen.
Existierende Corpora (wie z.B. die von [MS04, Sa07]) erfüllen diese Bedingungen nur teilweise oder gar nicht, aus diesem Grunde wurde ein eigener Corpus erstellt. Die Genres Forum und Blog wurden nochmals aufgeteilt, da deren jeweilige Start- und Post- bzw. Thread-Seiten sich strukturell stark unterscheiden. Für einen Überblick über die resultierenden Genres siehe Tab. 1.

Tabelle 1: Überblick über die verschiedenen im Corpus enthaltenen Genres

Übergeordnetes Web Genre	(Sub-)Genre	Beschreibung	Anzahl Seiten
Blog	Blog-Startseite	Typischerweise mehrere Posts (Teaser)	200
	Blog-Postseite	Einzelner Post mit Kommentaren	200
Forum	Forum-Startseite	Überblick über Themen und Threads	200
	Forum-Threadseite	Ein Thread mit mehreren Posts	200
Wiki	Wikiseite	Wikiseite mit einem Artikel	200
Outlier	Outlier	Unterschiedliche, heterogene Seiten	347
Total			**1347**

Die Seiten wurden aus unterschiedlichen Quellen randomisiert gewonnen und manuell überprüft. Schließlich ergibt sich ein Corpus aus 1347 Seiten mit sechs verschiedenen Genres. Das Genre Outlier wurde manuell aus einer Sammlung von Genres aus verwandten Ansätzen zusammengestellt und beinhaltet eine sehr heterogene Menge von Genres wie Download-Seiten, Suchergebnisseiten, Nachrichtenseiten und Lyrik. Diese Genres sind strukturell sehr unterschiedlich und repräsentieren alle Web Genres, die nicht in unserem Corpus gemessen werden sollen.

5 Evaluation und Ergebnisse

Für die Durchführung der Evaluation verwendeten wir das Weka Machine Learning Toolkit [WF05]. Für die Klassifizierung wurde eine Support Vector Machine eingesetzt, alle Ergebnisse wurden mit 10-fold Cross-Validation validiert, was bedeutet, dass der Corpus in zehn Teile partitioniert wird, von denen jeweils neun als Trainingsdaten benutzt werden und ein Teil als Testdaten. Dies erhöht die Validität der Ergebnisse durch einen sauberen Train/Test-Split.
In Tabelle 2 ist die Konfusionsmatrix des Ergebnisses der Evaluation zu sehen, die die richtig und falsch klassifizierten Instanzen sowie deren Verwechslung nach Genre abbildet. Wir erzielen mit unserem Ansatz eine Erkennungsgenauigkeit von 86%, was in Anbetracht der Tatsache, dass keine linguistischen Features erhoben werden und Outlier betrachtet werden, ein sehr gutes Ergebnis ist. Auffällig ist die Verwechslungsgefahr zwischen Blog-Start- und -Postseiten. Dies liegt daran, dass sie untereinander strukturell sehr ähnlich sind. Weiterhin scheint die Unterscheidung zwischen Blog-Startseiten und Outliers ebenfalls eine Fehlerquelle zu sein. Wir glauben, dass dies eine Folge der Heterogenität des „Genres“ Outlier in sich ist, d.h. Individuen dieses Genres folgen keinem einheitlichen Schema, sondern sind strukturell und inhaltlich sehr unterschiedlich. Daher besteht hier bei manchen Individuen eine Ähnlichkeit mit Blog-Startseiten, die selbst ein sehr heterogenes Genre zu sein scheinen.

Tabelle 2: Die Konfusionsmatrix der Evaluationsergebnisse unseres Ansatzes

a	**b**	**c**	**d**	**e**	**f**	← klassifiziert als
155	8	1	0	0	36	**a** = Blog-Startseite
18	173	0	1	0	9	**b** = Blog-Postseite
2	0	169	0	1	27	**c** = Wikiseite
1	0	0	173	3	23	**d** = Forum-Startseite
0	1	1	5	179	14	**e** = Forum-Threadseite
11	5	4	4	2	321	**f** = Outlier
0.83	0.93	0.97	0.95	0.97	0.75	**Precision**
0.76	0.86	0.85	0.87	0.89	0.93	**Recall**
1170 von 1347 korrekt klassifizierte Seiten (entspricht 86.8%)						

Unter den 25 aussagekräftigsten Features, die wir mittels Information Gain ermittelten, sind Häufigkeiten von HTML-Elementen (z.B. div und ul), die Pfad-Tiefe von URLs, der Grad der internen und externen Verlinkung sowie zwei unserer Pattern-Features (durchschnittliche Größe der Patterns und das Verhältnis von Patterns zu Text, der nicht in Patterns enthalten ist).
Diese Ergebnisse zeigen, dass eine Web-Genre-Erkennung ohne Einbeziehung von sprachlichen Features möglich ist und für unsere fokussierten Web-Genres mit 86% Erkennungsgenauigkeit für den Einsatz ausreichend zuverlässig ist.

6 Integration der Web-Genre-Erkennung in ELWMS.KOM

Die Erkennung des Web-Genres erfolgt bei der Persistierung der Webressourcen. Da der Client des ELWMS.KOM-Frameworks in ein Add-On für den Internetbrowser Firefox eingebettet ist , kann die Web-Genre-Erkennung sofort bei der Einordnung in das Wissensnetz den Quellcode analysieren und das Genre klassifizieren. Dieses Vorgehen ist insofern wichtig, als häufig keine vollständigen Webseiten, sondern nur Fragmente mit den relevanten Informationen im Wissensnetz gespeichert werden. Dies bedeutet, dass somit oft der Kontext verloren geht, der benötigt wird, um das Genre später zu ermitteln.

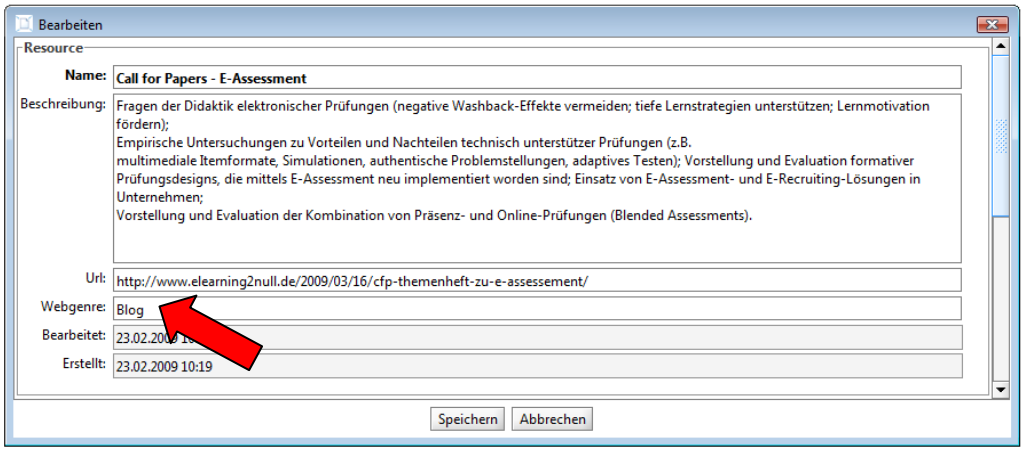

Abb. 4 Das Web-Genre eines Blogposts wurde automatisch erkannt und als Attribut für die Webressource eingetragen (Ausschnitt eines Screenshots)

Das ermittelte Genre wird automatisch in einem Metadatenfeld eingetragen und kann – sollte das ermittelte Genre nicht korrekt sein – vom Benutzer geändert werden (siehe Abb. 4). Es ist dem Benutzer auch durchaus möglich, ein eigenes Genre einzutragen, wobei dies keine Relevanz für zukünftige Klassifizierungen hat. Für das Retrieval kann diese Information bei einer Suche genutzt werden, indem zu Suchbegriffen zusätzlich eine Web-Genre-Angabe gemacht wird.

7 Zusammenfassung und Ausblick

In dieser Arbeit wurde gezeigt, dass eine automatische Web-Genre-Erkennung für Retrieval in (oft sehr umfangreichen) persönlichen und Community-Wissensnetzen sinnvoll ist und gute Ergebnisse zeigt. Der gewählte Ansatz hat gegenüber verwandten Arbeiten einige Vorteile wie Sprachunabhängigkeit und das Einbeziehen von Outliers. Eine Evaluation wurde präsentiert und die Integration in ELWMS.KOM aufgezeigt.
Die Ergebnisse der Web-Genre-Erkennung sind eine Grundlage für weitere Anwendungsszenarien: so kann das Genre als Grundlage zu einer semantischen Erkennung einer Auswahl dienen (z.B. ist der persistierte Inhalt Teil des Blogposts oder eines Kommentars?), die logische Identität einer Seite kann festgestellt werden (z.B. verschiedene Foreneinträge gehören logisch zusammen, obwohl sie aufgrund der Länge der Diskussion auf mehrere HTML-Seiten aufgeteilt wurden) oder ein Mapping der Wissensnetze kann aufgrund der Genre-Metadaten durchgeführt werden.
Diese Szenarien werden wir für unser Anwendungsszenario von persönlichen und Community-Wissensnetzen analysieren und in weiteren Arbeiten adressieren.

Literaturverzeichnis

[BS08a] Böhnstedt, D.; Scholl, P.; Benz, B.; Christoph Rensing; Steinmetz, R.; Schmitz, B.: Einsatz persönlicher Wissensnetze im Ressourcen-basierten Lernen DeLFI 2008: 6. e Learning Fachtagung Informatik, Lecture Notes in Informatic, 2008, S. 113-124.

[BS08b] Böhnstedt, D.; Scholl, P.; Rensing, C.; Steinmetz, R.: ELWMS.KOM – Typisiertes Tagging in persönlichen Wissensnetzen Workshop Proceedings der Tagungen Mensch & Computer 2008, DeLFI 2008 und Cognitive Design 2008, Logos Verlag, 2008, S. 330 331.

[DS08] Domínguez García, R.; Scholl, P.; Böhnstedt, D.; Rensing, C.; Steinmetz, R.: Towards an Automatic Web Genre Classification. Technical Report, Multimedia Kommunikation - Technische Universität Darmstadt, 2008.

[ER06] Elgersma, E.; de Rijke, M.: Learning to Recognize Blogs: A Preliminary Exploration. In: EACL 2006 Workshop on New Text – Wiki and Blogs and Other Dynamic Text Sources, 2006, S.24-30.

[MS04] Meyer zu Eissen, S. & Stein, B.: Genre Classification of Web Pages – User Study and Feasibility Analysis. In KI 2004: Advances in Artificial Intelligence, Springer Berlin / Heidelberg, 2004, S. 256-269.

[NP05] Nejdl, W. & Paiu, R.: I know I stored it somewhere - Contextual Information and Ranking on our Desktop. 8th International Workshop of the EU DELOS Network of Excellence on Future Digital Library Management Systems, 2005.

[Sa07] Santini, M.: Automatic Identification of Genre in Web Pages. University of Brighton, 2007.

[Va00] Vakkari, P.: Relevance and contributing information types of searched documents in task performance, Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval, ACM New York, NY, USA, 2000, S. 2-9.

[WF05] Witten, I.H.; Frank, E.: Data Mining: Practical machine learning tools and techniques, 2. Edition, Morgan Kaufmann, San Francisco, 2005.

[YS97] Yates, S.; Sumner, T.: Digital genres and the new burden of fixity. In: Hawaiian International Conference on System Sciences (HICSS 30), Hawaii, IEEE Computer Press, 1997.