Aspekte der Kategorisierung von Webseiten

Matthias Dehmer, Alexander Mehler und Rüdiger Gleim

Technische Universität Darmstadt
64289 Darmstadt
{dehmer | gleim}@informatik.tu-darmstadt.de

Universität Bielefeld
33501 Bielefeld
Alexander.Mehler@uni-bielefeld.de

Abstract: Im Zuge der Web-basierten Kommunikation tritt die Frage auf, inwiefern Webpages zum Zwecke ihrer inhaltsorientierten Filterung kategorisiert werden können. Diese Studie untersucht zwei Phänomene, welche die Bedingung der Möglichkeit einer solchen Kategorisierung betreffen (siehe [6]): Mit dem Begriff der *funktionalen Äquivalenz* beziehen wir uns auf das Phänomen, dass dieselbe Funktions- oder Inhaltskategorie durch völlig verschiedene Bausteine Web-basierter Dokumente manifestiert werden kann. Mit dem Begriff des *Polymorphie* beziehen wir uns auf das Phänomen, dass dasselbe Dokument zugleich mehrere Funktions- oder Inhaltskategorien manifestieren kann. Die zentrale Hypothese lautet, dass beide Phänomene für *Web-basierte Hypertextstrukturen* charakteristisch sind. Ist dies der Fall, so kann die automatische Kategorisierung von Hypertexten [2, 10] nicht mehr als eindeutige Zuordnung verstanden werden, bei der einem Dokument genau eine Kategorie zugeordnet wird. In diesem Sinne thematisiert das Papier die Frage nach der adäquaten Modellierung multimedialer Dokumente.

1 Einführung in die Problematik

Die Aufgabe der automatischen Textkategorisierung [4] besteht darin, textuelle Einheiten den Kategorien eines vorher definierten Kategoriensystems zuzuordnen. Gegenstand der automatischen Klassifikation von Hypertextstrukturen ist es, analog zur automatischen Textkategorisierung, hypertextuelle Einheiten (z.B. Webpages) auf eine vorgegebene Menge von Kategorien abzubilden. Maschinelle Lernverfahren der Textkategorisierung lassen sich in mehrere Gruppen unterteilen, wobei unter anderem *Vektorraum-basierte Verfahren* Anwendung finden. Für unser spezielles Kategorisierungsproblem wählen wir einen wichtigen Vertreter aus dieser Gruppe aus: die *Support Vector Machines* [1, 8]. Dieses Verfahren beruht darauf, dass die Daten in einen hochdimensionalen *Merkmalsraum* projiziert werden und mit Hilfe von *Kernelfunktionen* [7] nichtlineare Separierungen der zu klassifizierenden Daten sehr effizient vorgenommen werden können.

Die Aussagekraft einer Hypertext-Kategorisierungsstudie hängt entscheidend davon ab,

dass der Kategorisierung eine gründliche Analyse des zu klassifizierenden Inhalts vorausgeht: (i) Der zu klassifizierende Inhalt muss klassifizierbar sein — er muss sich möglichst eindeutig einer bestimmten Kategorie aus dem Kategoriensystem zuordnen lassen. (ii) Das Kategoriensystem muss sinnvoll und repräsentativ gewählt sein.

In dieser Untersuchung werden Webpages aus dem Bereich akademischer Konferenzen als zu kategorisierende Objekte betrachtet, wobei funktional abgrenzbare Einheiten, wie Seiten für den CfP, Anmeldedaten, Unterkünfte oder die elektronische Anmeldung, automatisch kategorisiert werden sollen. Die Problematik dieser scheinbar einfachen Kategorisierungsaufgabe, wird unmittelbar anhand von Abbildung (1) deutlich. Sie veranschaulicht, dass dieselbe Funktions- oder Inhaltskategorie — hier *Calls for Participation* — auf derselben oder — funktional äquivalent — über verschiedene Seiten präsentiert werden kann, wobei in letzterem Fall die gewünschte Untergliederung von Partizipationsarten mittels Links erreicht wird. Dieses einfache Beispiel verweist auf ein Phänomen, das wir bei *Web-basierten Hypertextstrukturen* beobachtet haben und dessen systematisches Vorkommen einer unmittelbaren Kategorisierung im oben erläuterten Sinn entgegensteht (siehe [6]): Seite A in Abbildung (1) ist insofern kategorial mehrdeutig, als sie zugleich mehrere Unterarten des Call for Participation manifestiert. Sollen diese Unterarten in der Kategorisierung separiert werden, so ist das Beispiel notwendigerweise mehreren Kategorien zuzuordnen. Wir sprechen in diesem Fall von *Polymorphie*: Dasselbe Dokument setzt sich aus Ausdruckseinheiten zusammen, die verschiedene Kategorien manifestieren. Dass aber das Beispiel B aus Abbildung (1) überhaupt als funktional (partiell) äquivalent zu Beispiel A gelten kann, liegt daran, dass verschiedene Komponenten von Webpages ähnliche Funktionen übernehmen können: Links sind in diesem Beispiel beispielsweise durch eine stärkere Dokumentuntergliederung ersetzbar. In diesem Fall sprechen wir von *funktionaler Äquivalenz* beider Ausdrucksmittel. Die Hypothese lautet nun: Wenn funktionale Äquivalenz und Polymorphie charakteristische Eigenschaften Web-basierter Strukturen sind, dann können Webpages nicht länger als eindeutig kategorisierbare Einheiten gelten, da *polymorphe* Webpages mehrere Kategorien instanziieren. Somit wird ein relationaler Zusammenhang von Hypertextstrukturtypen und Kategorien erwartet. Dass wir von Polymorphie/funktionaler Äquivalenz anstelle von Polysemie/Synonymie sprechen, liegt nicht nur daran, dass letztere Termini primär auf lexikalischer Ebene Anwendung finden, sondern auch daran, dass der Polymorphiebegriff in der Linguistik nicht allein auf die Mehrdeutigkeit, sondern auch auf die strukturelle Variabilität von Zeichen fokussiert.

2 Das Testkorpus

Um letztere Hypothese zu überprüfen, betrachten wir ein indirektes Experiment: die Kategorisierung eines annahmegemäß hochgradig strukturierten Bereichs des Webs, und zwar Konferenz-Homepages. Diese können als hochgradig strukturiert gelten, da sie rekurrente Funktionen auf erwartbare Art und Weise zu bedienen haben: Anmeldung, Programmübersicht, Hinweise über Unterkünfte, etc. Diesem Gedanken folgend wählen wir englischsprachige Konferenz-Websites im Bereich „Computer Science“und „Mathematics“. Um nun das Testkorpus bestehend aus 13.481 Webpages zu konstruieren, er-

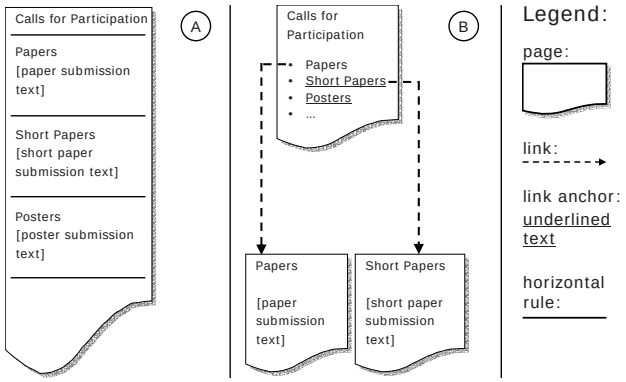

Abbildung 1: Schematische Darstellung zweier funktional äquivalenter Präsentationen mittels (A) einer Liste und (B) eines *compound document* bestehend aus mehreren Webpages.

stellten wir mit einer Java-Applikation ausgehend von Konferenz-Kalender-Webpages die entsprechende Menge von Konferenzlinks. Basierend auf dieser Menge von Links extrahierten wir mit einem von uns für die strukturelle Analyse von Hypertextstrukturen entwickelten Tools `HyGraph` die jeweiligen Websites und transformierten sie in eine auf dem Graphenaustauschformat `GXL` [9] basierende Graphrepräsentation. Um schließlich die Kategorisierung der Webpages mit der Support Vector Machine-Bibliothek `LibSVM` [3] vorzunehmen, wurden für die Webpages Tokenvektor-Darstellungen erzeugt.

3 Evaluierung

Wir definieren zunächst formal unser Kategorisierungsproblem und beschreiben im folgenden die Schritte der Evaluierung. Es sei $K := \{K_1, K_2, \ldots, K_{|K|}\}$ eine Menge von Kategorien, $T := \{T_1, T_2, \ldots, T_{|K|}\}$ eine Menge von Trainingsmengen und U, $|U| < \infty$, die Menge der noch nicht kategorisierten Webpages. In dieser Studie besteht die Kategorisierungsaufgabe darin, den extrahierten Webpages $u \in U$ Kategorien zuzuordnen, die basierend auf dem textuellen Inhalt der Webpages den funktionalen Typ oder die Semantik widerspiegeln (z.B. *list of accepted papers*). *Polymorphie* und *funktionale Mehrdeutigkeit* drückt sich nun darin aus, dass keine eindeutige Zuordnung $U \ni u \mapsto K_i \in K$ existiert — die Zuordnung zwischen den Webpages und den Kategorien entartet zu einer Relation:

$$\mathcal{R} \subseteq U \times K, \mathcal{R} := \{(u, K_i) | U \ni u \text{ gehört zur Kategorie } K_i \in K\}.$$

Die Menge der Kategorien ist in unserem Experiment wie folgt definiert: $K = \{$*submission and author instructions, call for papers, important dates, committees, accepted papers, topics and general information, program, travel and accommodation, venue, invited speakers, registration, sponsors, workshops*$\}$, so dass $|K| = 13$. Für jedes $K_i, 1 \leq i \leq |K|$ definieren wir eine binäre Klassifikation (*one against all*) und müssen somit für jede Kategorie eine Trainingsmenge $T_i, 1 \leq i \leq |K|$ und den dazugehörigen optimalen Para-

K_i	precision	recall	accuracy
K_1	29,1%	99,0%	70,8%
K_2	41,6%	99,0%	82,5%
K_3	41,2%	99,0%	90,4%
K_4	50,0%	99,2%	88,2%
K_5	66,6%	99,0%	72,1%
K_6	35,0%	99,1%	90,4%
K_7	25,5%	66,0%	68,4%
K_8	50,0%	99,2%	80,3%
K_9	32,0%	99,0%	66,3%
K_{10}	25,0%	99,0%	80,1%
K_{11}	46,1%	99,0%	71,3%
K_{12}	41,6%	99,0%	82,9%
K_{13}	52,1%	99,2%	94,1%

Abbildung 2: Ergebnisse der Performanzmessung.

K_i	# matchings	U_i
K_1	2107	0,10
K_2	2661	0,05
K_3	1992	0,05
K_4	1546	0,24
K_5	3846	0,02
K_6	3616	0,02
K_7	2716	0,14
K_8	2245	0,03
K_9	3045	0,02
K_{10}	2206	0,01
K_{11}	3339	0,03
K_{12}	4627	0,03
K_{13}	1141	0,02

Abbildung 3: Zur Messung von Polymorphie.

metervektor bestimmen. Um die Trainingsmenge T_i basierend auf der Kategorie K_i zu konstruieren, sei $l^+_{K_i}$ die Anzahl der positiven Trainingsbeispiele für Kategorie K_i und $l^-_{K_1}, l^-_{K_2}, \ldots, l^-_{K_{i-1}}, l^-_{K_{i+1}}, \ldots, l^-_{K_{|K|}}$ die Anzahlen der zufällig aus der Menge aller Trainingsbeispiele gezogenen negativen Beispiele der verbleibenden Kategorien. Dabei sind $l^-_{K_1}, l^-_{K_2}, \ldots, l^-_{K_{i-1}}, l^-_{K_{i+1}}, \ldots, l^-_{K_{|K|}}$ annähernd gleich groß gewählt und es gilt für die Indexmenge $I_{-i} := \{1, 2, \ldots, i-1, i+1, \ldots, |K|\}$ die Bedingung $\sum_{j \in I_{-i}} l^-_{K_j} = l^+_{K_i}$. Für die Kategorisierungsaufgabe verwenden wir den SVM-Typ C-SVM der `LibSVM`-Bibliothek und die RBF-*Kernelfunktion* vom Typ $K(u, v) := e^{-\gamma \|u-v\|^2}$. Um die optimalen Parametervektoren der Form (C, γ) für die oben konstruierten Trainingsmengen zu bestimmen, führen wir eine Suche im Parameterraum $P := \{(C, \gamma) | C = 2^g, \gamma = 2^s, g \in \{-4, 0, 4, \ldots, 20\}, s \in \{-16, -12, -8, \ldots, 8\}\}$ durch, kombiniert mit einer *5-fold cross validation*. Dabei wurden für jedes K_i diejenigen Parametervektoren ausgewählt, welche den Fehler der *Cross Validation* bezogen auf die Trainingsmenge T_i minimieren.

Mit Hilfe der aus dem *Information Retrieval* bekannten Performanzmaße *Recall*, *Precision* und *Accuracy* führten wir eine Performanzmessung der SVM-Kategorisierung durch. Tabelle (2) fasst das Ergebnis zusammen: hohe *Recall*- und niedrige *Precision*-Werte. Das bedeutet für die Kategorienmenge $\hat{K} := K \backslash \{K_7\}$, dass Webpages $u \in U$ fast immer den Kategorien zugeordnet werden, denen sie angehören, darüber hinaus vielfach aber auch solchen, denen sie nicht angehören. Es handelt sich also um eine hochgradig fehlerhafte Kategorisierung, und zwar im Rahmen des gewählten Vektorraummodells und SVM-Klassifikators. Dieser negative Eindruck wird durch den Eindeutigkeitskoeffizienten $U_i \in [0, 1]$ bestätigt, der angibt, wieviele der Testseiten, die einer Kategorie K_i zugeordnet wurden, ausschließlich dieser Kategorie zugeordnet werden. Die Koeffizienten U_i geben also einen Eindruck von der Trennschärfe der betrachteten Kategorienmenge. Dazu gelte

$[|K_i(u)|] = 1$ gdw. u der Kategorie K_i angehört. Es ist dann:

$$U_i := \frac{|\{u \in U \mid K_i(u) \wedge \neg(K_1(u) \vee \ldots \vee K_{i-1}(u) \vee K_{i+1}(u) \vee \ldots \vee K_{|K|}(u))\}|}{|\{u \in U \mid K_i(u)\}|}.$$

Tabelle (3) demonstriert die extrem geringe Trennschärfe der Kategorienmenge, was darauf hinweist, dass entweder die falschen Merkmale ausgewählt wurden oder der falsche Klassifikator oder — so unsere noch weiter zu untermauernde Interpretation — die betrachteten Webpages systematisch durch Polymorphie gekennzeichnet sind.

4 Zusammenfassung und Ausblick

Das Paper berichtete von einer indirekten Messung der Wirksamkeit von Polymorphie im Bereich Web-basierter Dokumente. Ist dieses Phänomen auch in weiteren Experimenten vergleichbarer Textklassen nachweisbar, so bedeutet das, dass das Standardrepräsentationsmodell der Kategorisierung, das Vektorraummodell also, unzureichend ist, da es die Struktur von Dokumenten außer Acht lässt. Der Messbarmachung dieser Art struktureller Polymorphie werden wir zukünftige Kategorisierungsexperimente widmen.

Literatur

[1] Cristianini N., Shawe-Taylor J.: *An Introduction to Support Vector Machines*, Cambridge University Press, Cambridge, UK, 2000

[2] Fürnkranz J.: *Hyperlink Ensembles: A Case Study in Hypertext Classification,*Technical Report OEFAI-T-2001-30, 2001

[3] Hsu C.-W., Chang C.-C., Lin C.-J.: *A practical guide to SVM classification*, Technical report, Department of Computer Science and Information Technology, National Taiwan University, 2003

[4] Joachims T.: *Learning to classify text using support vector machines*, Kluwer, Boston, 2002.

[5] De Lara E., Wallach D. S., Zwaenepoel W.: *A Characterization of Compound Documents on the Web,* Rice Computer Science Technical Report T99-351, 1999

[6] Mehler A., Dehmer M., Gleim R.: *Towards Logical Hypertext Structure. A Graph-Theoretic Perspective*, erscheint in: Proc. of I2CS '04. Berlin/New York: Springer.

[7] Schölkopf B., Müller K. R., Smola A.J: *Lernen mit Kernen. Support-Vektor-Methoden zur Analyse hochdimensionaler Daten*, Informatik Forsch. Entw., Vol. 14, 1999, 154-163

[8] Vapnik V.: *The nature of Statistical Learning Theory*, Springer Verlag, 1995

[9] Winter A., Kullbach B., Riedinger V.: *An overview of the GXL graph exchange language*, In Software Visualization, Springer Berlin/Heidelberg, 2002, 324-336

[10] Yang Y., Slattery S., Ghani R.: *A Study of Approaches to Hypertext Categorization*, Journal of Intelligent Information Systems, Vol. 18(2-3), 2002, 219-241