

# Einsatz Maschinelles Lernverfahren in der Dokumentklassifikation

– Eine Standortbestimmung –

**Markus Junker** und **Andreas Abecker**  
Deutsches Forschungszentrum für Künstliche Intelligenz  
– DFKI GmbH –  
Postfach 2080, 67608 Kaiserslautern  
e-mail: {junker|abecker}@dfki.uni-kl.de

## 1 Problemstellung

Das Information Retrieval (IR) beschäftigt sich mit dem inhaltsorientierten Zugriff auf große Dokumentsammlungen. Eine wichtige Aufgabe dabei ist die Zuweisung eines (Text-)Dokuments zu bestimmten, über ihren Inhalt definierten Dokumentklassen (Dokumentklassifikation). Die durch Dokumentklassifikationssysteme vorgenommenen Klassenzuweisungen dienen zum Archivieren oder zum inhaltsbasierten Weiterleiten von Dokumenten an bestimmte Adressaten. Innerhalb der Dokumentanalyse werden die Klassifikationsergebnisse auch zur gezielten Ansteuerung bestimmter Analysealgorithmen verwendet (vgl. [Dengel et al. 95]).

In traditionellen, *wissensbasierten* Dokumentklassifikationssystemen wird die Einordnung der Dokumente durch handkodierte Regeln vorgenommen, welche auf das Vorkommen bestimmter Wörter und Wortpattern testen. Zur Bildung von Pattern kann man z.B. das Auftreten zweier Wörter in einem bestimmten Maximalabstand fordern oder das Auftreten bestimmter Wörter verbieten. Das bekannteste derartige System ist bei der Nachrichtenagentur Reuters mit großem Erfolg im Einsatz (*CONSTRUE* [Hayes et al. 88, Hayes et al. 90]). Im Gegensatz zu wissensbasierten Systemen sollen *lernende* Dokumentklassifikationssysteme sich das zur Klassifikation neuer Dokumente benötigte Wissen selbständig aus vorklassifizierten Trainingsdokumenten aneignen. Hierdurch soll der teilweise erhebliche Aufwand zur Erstellung der Wissensbasis minimiert werden.

Die Dokumentklassifikation ist auch eine wesentliche Aufgabe in unserem Dokumentanalyse-system *OfficeMAID*, welches sich exemplarisch mit der Analyse deutscher Geschäftsbriefe befaßt [Dengel et al. 94, Dengel et al. 95]. Grundlage für die Dokumentklassifikation in *OfficeMAID* bilden die Ergebnisse einer optischen Zeichenerkennung (OCR) auf eingescannten Dokumenten. Im Rahmen von *OfficeMAID* wurden bereits zahlreiche Erfahrungen in der wissensbasierten Dokumentklassifikation gesammelt [Hoch 94, Wenzel et al. 95, Wenzel et al. 96]. Wir untersuchen nun die Adaption von Verfahren des maschinellen Lernens für unser Problem. Im Rahmen dieses Artikels gehen wir zuerst kurz auf bestehende Ansätze zum Einsatz von Lernverfahren in der Dokumentklassifikation ein. Eher bescheidene Erfolge in bisherigen Untersuchungen zeigen, daß die naive Verwendung von Standard-Lernverfahren auf einfachen Dokumentrepräsentationen zu kurz greift. Stattdessen sind sorgfältige Analysen der verfügbaren Informationen und der spezifischen Charakteristika des Lernproblems erforderlich, um Anforderungen an (ggf. neu zu entwickelnde)

Lernverfahren für die Dokumentklassifikation zu spezifizieren. Für unsere Geschäftsbriefdomäne haben wir damit begonnen und präsentieren im folgenden erste Überlegungen zur Gestaltung spezifischer Klassifikationsverfahren.

## 2 Maschinelles Lernen in der Dokumentklassifikation

Abbildung 1 veranschaulicht das typische Szenario beim Einsatz von Lernverfahren in der Dokumentklassifikation. Im dargestellten Modell wird zwischen Lern- und Anwendungsphase eines Dokumentklassifikationssystems unterschieden. In der Lernphase werden zunächst aus der Trainingsmenge geeignete Attribute (Features) erzeugt (1). Mit Hilfe dieser Features werden die Dokumente der Trainingsmenge in die korrespondierenden Featurredarstellungen transformiert (2). Schließlich werden Lernverfahren benutzt, um aus den in dieser Repräsentation vorliegenden Dokumenten einen geeigneten Klassifikator zu konstruieren (3). Zur Klassifikation eines neuen Dokuments in der Anwendungsphase des Systems muß dieses in die korrespondierende Featurredarstellung transformiert (4) und hierauf der zuvor erzeugte Klassifikator angewendet werden (5).

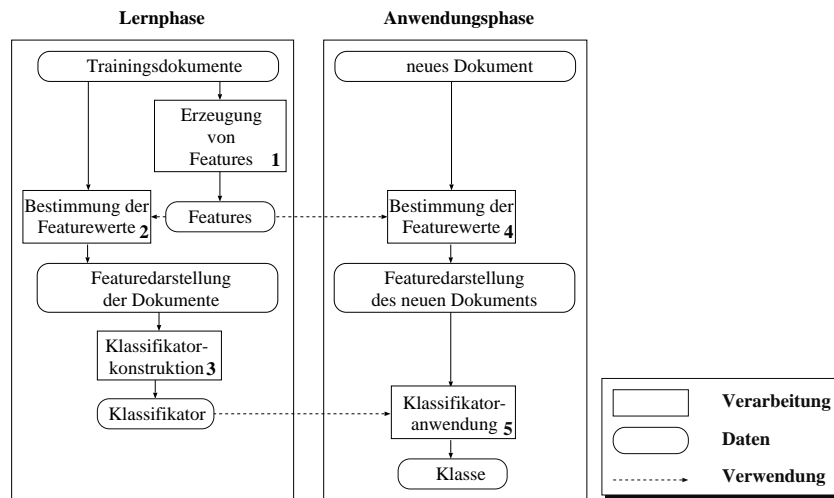


Abbildung 1: Modell für lernfähige Dokumentklassifikationssysteme

Im Bereich lernender Dokumentklassifikationssysteme gibt es zahlreiche aus dem klassischen IR adaptierte Ansätze (beruhend auf Nearest-Neighbor-Verfahren, naive Bayes-Verfahren, Prototyp-Klassifikation oder Neuronalen Netzen—siehe z.B. [Lewis 92a, Ittner et al. 95] und viele der in [Harman 95] vorgestellten Verfahren). Neben solchen Ansätzen besteht auch großes Interesse an der Erschließung symbolischer Verfahren des maschinellen Lernens für die Dokumentklassifikation [Apté et al. 94, Lewis et al. 94, Tong et al. 94, Goldberg 95]. Beispiele für an der Dokumentklassifikation erprobte symbolische Lernverfahren sind SWAP [Weiss et al. 93], C4.5 [Quinlan 93] und CART [Breiman et al. 84]. Teilweise wurde auch versucht, durch Konstruktion eigener Lernverfahren bessere Ergebnisse zu erzielen [Goldberg 95, Finch 95].

In der Regel werden zur Erzeugung der Features alle in den Dokumenten der Trainingsmenge enthaltenen Wörter herangezogen. Für jedes auftretende Wort wird ein Feature generiert, dessen Wert für ein Dokument sich aus der Häufigkeit seines Auftretens im betreffenden Dokument er-

gibt. Oft wird auch eine Reduktion der Wörter auf ihre Stammform durchgeführt, Stoppwörter<sup>1</sup> werden entfernt sowie zusätzlich Paare benachbarter Wörter als Features eingeführt. Über eine solche einfache Dokumentrepräsentation hinausgehende Techniken wie die Aufnahme allgemeinerer Wortpattern wurden bisher nur vereinzelt und größtenteils mit bescheidenem Erfolg genutzt [Lewis 92b].

Für maschinelle Lernverfahren stellt sich die Dokumentklassifikation recht unfreundlich dar (vgl. [Cohen 95, Lewis 92a]): Man hat sehr viele Features, von denen für ein konkretes Dokument aber nur sehr wenige relevant sind. Andererseits gibt es auch Features, die nur für sehr wenige Dokumente relevant sind. Bei der gleichzeitigen Aufnahme von Wortpattern und der in ihnen enthaltenen Einzelwörter in die Featuredarstellung entstehen starken Abhängigkeiten, die Verfahren wie naive Bayes oder einfache Nearest Neighbor-Methoden nicht ohne weiteres angemessen verarbeiten können. Häufig hat man überlappende Klassen, die sich mit den vorliegenden Features eher beschreiben als von anderen Klassen separieren lassen. Deshalb lernt man i.a. binäre Klassifikatoren, die nur eine bestimmte Klasse erkennen sollen. Für das Trainieren eines solchen Klassifikators hat man dann aber meist nur sehr wenige positive Beispiele im Vergleich zur Anzahl der negativen. Im Unterschied zu gängigen Lernansätzen ist das Lernziel bei der Dokumentklassifikation nicht zwingend eine geringe Fehlerquote, sondern häufig über eine Gewichtung von Precision und Recall definiert.

### 3 Mögliche Verbesserungsansätze

Unsere Erfahrungen im Bereich der Klassifikation von Geschäftsbriefen zeigen, daß für einen erfolgreichen Einsatz maschineller Lernverfahren die Repräsentationssprache für Klassifikatoren erweitert werden sollte. Dafür sehen wir Verbesserungspotential in mindestens zwei Bereichen. Einerseits kann man in vielen Fällen über weit mehr Information verfügen, als die einfachen, auf dem bloßen Text basierenden Verfahren nutzen. Abschnitt 3.1 führt einige solcher Zusatzinformationen und ihre mögliche Nutzbarmachung an. Andererseits sollte die Darstellung auch von allgemeineren Wortpattern—wie in der wissensbasierten Klassifikation üblich—möglich sein. Dazu schlagen wir in Abschnitt 3.2 die Konstruktion von Wortpattern als Spezialisierungsoperator beim Regellernen vor. Regellernverfahren bieten sich aus verschiedenen Gründen an: sie stellen die Spezialisierung von Hypothesen explizit dar, sind relativ robust gegenüber Abhängigkeiten zwischen Features, liefern dem Benutzer verständliche Resultate und lassen sich in ihren Ergebnissen gut mit manuell erstellten wissensbasierten Klassifikatoren vergleichen.

#### 3.1 Aufnahme struktureller Informationen in die Dokumentrepräsentation

Sollen, wie in unserem Fall, eingescannte und durch OCR bearbeitete Dokumente klassifiziert werden, so kann man häufig auf weit mehr Informationen als nur die erkannten Wörter zurückgreifen. Hierzu gehören beispielsweise Fontinformationen wie Größe oder Schrifttyp eines Wortes und seine geometrische Position auf dem Dokument. Bisherige Ergebnisse auf unseren Geschäftsbriefen zeigen, daß durch die Hinzunahme solcher Informationen die Klassifikationsgüte deutlich gesteigert werden kann. Beispielsweise tritt das Wort **Rechnung** nicht nur in Rechnungen, sondern auch in anderen Dokumentklassen auf. Auffällig ist jedoch, daß es in Rechnungen häufig im oberen Drittel des Dokumentes zu finden ist und durch Größe bzw. Font hervorgehoben wird. Die

---

<sup>1</sup>Als Stoppwörter werden Wörter bezeichnet, die sicher nicht zur Bedeutung eines Textes beitragen (z.B. Artikel und Präpositionen).

übliche, flache Featuredarstellung ist unseres Erachtens nicht sonderlich geeignet zur Repräsentation derartiger Information. Hier bietet sich eher eine objektartige Darstellung an, wobei jedes Auftreten eines Wortes durch ein eigenes Objekt dargestellt wird. Dieses Objekt stellt zusätzliche Informationen über die entsprechende Ausprägung des Wortes, wie seine Position oder Fontgröße bereit. Wird bei der Regelkonstruktion festgestellt, daß das Auftreten eines Wortes ein gutes Diskriminierungsverhalten liefert, kann dieses Auftreten mithilfe von Anforderungen an seine Ausprägung weiter spezialisiert werden.

Auch in bereits in Textform vorliegenden Dokumenten lassen sich oftmals Zusatzinformationen zu Wörtern herleiten. Hierzu gehört beispielsweise, ob ein Wort in einem bestimmten Feld steht (z.B. im Subject-Feld einer Email) oder Teil einer Überschrift ist (erkennbar an Formatierungsanweisungen oder Absetzung). Derartige Informationen könnten ebenfalls wie beschrieben repräsentiert und genutzt werden. Die Umsetzung des vorgeschlagenen Konzepts kann in Anlehnung an bekannte Verfahren realisiert werden, die sich mit hierarchischen Attributen beschäftigen (siehe z.B. [Breiman et al. 84]).

### 3.2 Integration der Wortpatternkonstruktion in das Lernverfahren

Neben den aufgeführten strukturellen Informationen, kann auch die Berücksichtigung des Wortkontextes zu besseren Klassifikationsergebnissen führen. Aus der Verwendung des Wortes *Rechnung* innerhalb bestimmter Wortpattern ist oft abzuleiten, ob es sich tatsächlich um ein Dokument der Klasse *Rechnung* handelt oder nicht. Findet man in einem Dokument das Pattern *Rechnung vom*, so wird es sich mit hoher Wahrscheinlichkeit nicht um ein Rechnung, sondern lediglich um ein auf eine Rechnung bezugnehmendes Schreiben handeln.

In den bisherigen Ansätzen wurden zumeist nur sehr einfache Wortpattern berücksichtigt, welche vor der eigentlichen Lernphase berechnet und in die Featuredarstellung aufgenommen werden. Uns erscheint es sinnvoller, zunächst mit einer einfachen, lediglich auf Einzelwörtern beruhenden Featuredarstellung zu beginnen und die Konstruktion geeigneter Wortpattern in die Regelkonstruktion einzubauen. Analog zur bei Regellernverfahren üblichen Spezialisierung durch konjunktive Verknüpfung mit Featureausprägungen, sollen dabei Einzelwörter zu besser diskriminierenden Wortpattern spezialisiert werden. Im Gegensatz zur bisherigen Praxis erlaubt dies eine zielorientiertere Konstruktion von Wortpattern, welche auch allgemeinere Wortpattern berücksichtigen könnte. Wir planen im Rahmen eines Regellernverfahrens zwei alternative Strategien empirisch zu evaluieren:

1. Sobald ein einzelnes Wort  $w$  sich für das Klassifikationsproblem als ausreichend relevant erweist, werden alle in der Trainingsmenge in einem Maximalabstand  $n$  vorkommenden Wörter  $w_1 \dots w_k$  zur Bildung von Wortpattern  $ww_1 \dots ww_k$  herangezogen. Der Wert dieser Pattern ergibt sich aus dem Abstand der beteiligten Wörter.
2. Sobald sich das Zusammenauftreten zweier Wörter  $w_1$  und  $w_2$  innerhalb eines Dokuments als ausreichend relevant erweist, wird die Spezialisierung dieser Dokumenteigenschaft durch die Forderung eines Maximalabstandes  $n$  zwischen den betreffenden Wörtern geleistet. Hierzu wird ein neues Feature  $w_1w_2$  eingeführt, dessen Wert für ein Dokument sich wiederum aus dem Abstand der jeweiligen Wörter berechnet.

Insbesondere beim zweiten Verfahren ist es zur Suchraumbeschränkung sinnvoll, einen Maximalabstand  $n$  zwischen Wörtern festzulegen. Haben die beteiligten Wörter einen größeren

Abstand als  $n$ , so wird lediglich ihr Zusammenauftreten innerhalb eines Dokuments berücksichtigt. Die obige Darstellung ist insofern vereinfacht, als auch das Nichtauftreten von Wörtern zur Bildung von Pattern herangezogen werden kann. Weiterhin können auch bereits hergeleitete Wortpattern erneut spezialisiert werden.

Der Unterschied in den beiden Strategien besteht in den Voraussetzungen, unter welchen Wörter zu Pattern expandiert werden. Die erste Strategie versucht, bei Vorfinden eines ausreichend relevanten Wortes direkt, dieses zum Bestandteil eines noch aussagekräftigeren Wortpattern zu machen und zieht hierzu alle Wörter in der unmittelbaren Nachbarschaft heran. Dies ist mit einem erhöhten Suchaufwand zur Konstruktion von Pattern verbunden, kann sich aber dann positiv auswirken, wenn aussagekräftige Wortpattern schlecht diskriminierende Einzelwörter enthalten. Geht man davon aus, daß das gemeinsame Auftreten von an einem Wortpattern beteiligten Wörter bereits ein gutes Diskriminierungsverhalten liefert, so scheint die zweite Strategie angebracht. Man spart hierdurch im Bereich der Wortpatternkonstruktion Suchaufwand, welcher an anderer Stelle sinnvoller verwendet werden kann (beispielsweise zur Ansteuerung eines Wrapper-Algorithmus zur Justierung von Parametern des verwendeten Lernverfahrens).

## 4 Zusammenfassung

Wir haben die Dokumentklassifikation als relativ junges Gebiet für den Einsatz maschineller Lernverfahren beschrieben, das große praktische Relevanz besitzt. Bisherige Untersuchungen zum Lernen in der Dokumentklassifikation betrafen überwiegend die Ad-Hoc-Verwendung konventioneller Lernverfahren und lieferten teilweise bescheidene Ergebnisse. Wir betrachten die unzureichende Nutzung verfügbarer Informationen und die nicht genügend ausdrucksmächtige Hypothesensprache als mögliche Ursachen. Eingehende Untersuchungen der Domäne müssen die Parametrierung von Lernverfahren triggern und ggf. zu Verfahrensmodifikationen führen. Ähnliche Überlegungen führt Cohen in verschiedenen neueren Publikationen an, wo er z.B. den Einsatz empirischer ILP-Methoden und die Berücksichtigung von Wortkontextinformation vorschlägt [Cohen 95, Cohen 96]. Unsere eigenen Arbeiten in nächster Zeit betreffen die Implementierung eines modifizierten CN2-Verfahrens und die weitere Analyse unserer Geschäftsbriefdomäne.

## Literatur

- [Apté et al. 94] C. Apté, F. Damerau und S. Weiss. Towards Language Independent Automated Learning of Text Categorization Models. In: W. Croft und C.J. van Rijsbergen (Herausgeber), *Proceedings of the 17th Annual International ACM/SIGIR Conference on Research and Development in Information Retrieval*, Seite: 23–30, Dublin, Ireland, 3-6 July 1994.
- [Breiman et al. 84] L. Breiman, J.H. Friedman, R.A. Olshen und C.J. Stone. *Classification and Regression Trees*. Wadsworth International Group, 1984.
- [Cohen 95] W.W. Cohen. Text Categorization and Relational Learning. In: *Proceedings of the 12th International Conference on Machine Learning*, Seite: 124–132, Tahoe City, CA, USA, July 9-12 1995.
- [Cohen 96] W.W. Cohen. Context-Sensitive Learning Methods for Text Categorization. In: *Proceedings of the 19th Annual International ACM/SIGIR Conference on Research and Development in Information Retrieval (SIGIR 96)*, Zurich, Switzerland, August 18 - 22 1996. Accepted for publication.
- [Dengel et al. 94] A. Dengel, R. Bleisinger, F. Fein, R. Hoch, F. Hönes und M. Malburg. OfficeMAID – A System for Office Mail Analysis, Interpretation and Delivery. In: *Proceedings of the First International Workshop on Document Analysis Systems (DAS 94)*, Seite: 253–275, Kaiserslautern, Germany, October 1994.

- [Dengel et al. 95] A. Dengel, R. Bleisinger, F. Fein, R. Hoch, F. Hönes und M. Malburg. OfficeMAID – A System for Office Mail Analysis, Interpretation and Delivery. *Document Analysis Systems, Series in Machine Perception and Artificial Intelligence*, 14 Seite: 52–75, 1995.
- [Finch 95] S. Finch. Partial Orders for Document Representation: A new Methodology for Combining Document Features. In: E.A. Fox, P. Ingwersen und R. Fidel (Herausgeber), *Proceedings of the 18th Annual International ACM/SIGIR Conference on Research and Development in Information Retrieval (SIGIR 95)*, Seite: 264–272, Seattle, Washington, USA, 9-13 July 1995.
- [Goldberg 95] J.L. Goldberg. CDM: An Approach to Learning in Text Categorization. In: *6th International Conference on Tools with Artificial Intelligence*, Seite: 258–265, Washington, DC, USA, November 5-8 1995.
- [Harman 95] D. K. Harman (Herausgeber). *Proceedings of the Third Text REtrieval Conference (TREC-3)*. NIST Special Publication 500-225, 1995.
- [Hayes et al. 88] P.J. Hayes, L.E. Knecht und M.J. Cellio. A News Story Categorization System. In: *Proceedings of the 2rd Conference on Applied Natural Language Processing*, Seite: 9–17, Austin, TX, USA, February 9-12 1988.
- [Hayes et al. 90] P.J. Hayes, P.M. Anderson, I.B. Nirenburg und L.M. Schmandt. TCS: A Shell for Content-Based Text Categorization. In: *Proceedings of 6th Conference on Artificial Intelligence Applications*, Seite: 320–326, Santa Barbara, CA, USA, May 5-9 1990.
- [Hoch 94] R. Hoch. Using IR Techniques for Text Classification in Document Analysis. In: W. Croft und C.J. van Rijsbergen (Herausgeber), *Proceedings of the 17th Annual International ACM/SIGIR Conference on Research and Development in Information Retrieval*, Seite: 31–40, Dublin, Ireland, 3-6 July 1994.
- [Ittner et al. 95] D.J. Ittner, D.D. Lewis und D.D. Ahn. Text Categorization of Low Quality Images. In: *Proceedings of the 4th Annual Symposium on Document Analysis and Information Retrieval (SDAIR 95)*, Seite: 1–13, Las Vegas, NV, USA, 24–26 April 1995.
- [Lewis 92a] D. D. Lewis. *Representation and Learning in Information Retrieval*. Dissertation, Department of Computer Science, University of Massachusetts, 1992.
- [Lewis 92b] D.D. Lewis. An Evaluation of Phrasal and Clustered Representations on a Text Categorization Task. In: *Proceedings of the 15th Annual International ACM/SIGIR Conference on Research and Development in Information Retrieval*, Seite: 37–50, Copenhagen, Denmark, 21-24 June 1992.
- [Lewis et al. 94] D.D. Lewis und M. Ringuette. A Comparison of Two Learning Algorithms for Text Categorization. In: *Proceedings of the 3rd Symposium on Document Analysis and Information Retrieval (SDAIR 94)*, Seite: 81–93, Las Vegas, NV, USA, April 11-13 1994.
- [Quinlan 93] J.R. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers, Inc., 1993.
- [Tong et al. 94] R.M. Tong und L.A. Appelbaum. Machine Learning for Knowledge-Based Document Routing. In: D.K. Harman (Herausgeber), *2nd Text Retrieval Conference (TREC-2)*, Seite: 253–264, Gaithersburg, MD, USA, 31 August - 2 September 1994.
- [Weiss et al. 93] S.M. Weiss und N. Indurkha. Optimized Rule Induction. *IEEE Expert*, 8(6) Seite: 61–69, December 1993.
- [Wenzel et al. 95] C. Wenzel und R. Hoch. Text Categorization of Scanned Documents Applying a Rule-Based Approach. In: *Proceedings of the 4th Annual Symposium on Document Analysis and Information Retrieval (SDAIR 95)*, Seite: 333–346, Las Vegas, NV, USA, 24-26 April 1995.
- [Wenzel et al. 96] C. Wenzel, S. Baumann und Th. Jäger. Advances in Document Classification by Voting of Competitive Approaches. In: *International Association for Pattern Recognition Workshop on Document Analysis Systems (DAS 96)*, Malvern, Pennsylvania, USA, October 14-16 1996. Accepted for publication.