FAKULTÄT ELEKTROTECHNIK UND INFORMATIK
Hochschule Zittau/Görlitz

Datenanalyse für die automatische Klassifikation von Webseiten am Beispiel von Wertpapierdaten

BACHELORARBEIT

ANGEFERTIGT DURCH	Stefan Radusch
BETREUT DURCH	Prof. Dr.-Ing. G. Ringwelski
ANGEFERTIGT BEI	Deutsche Software Engineering & Research GmbH

Das Auffinden von relevanten Daten aus einer Internetrecherche kann automatisch durch Klassifikatoren geschehen. Diese Klassifikatoren müssen maschinelles Lernen anwenden, damit darauf folgend unbekannte Daten kategorisiert werden können. Die Frage, die sich allerdings stellt: Nach welchen Merkmalen muss ein Klassifikator lernen, um beim Klassifizieren möglichst wenig Fehler zu machen?

In dieser Arbeit wird ein Naïve-Bayes-Klassifikator verwendet, der mittels einfacher Textanalyse, Schlüsselwörtern, kontextsensitiven und semantischen Merkmalen Webseiten klassifiziert. Dabei wurde festgestellt, dass eine normale Textanalyse nicht für eine Klassifikation geeignet ist. Mit Schlüsselwörtern lässt sich eine Klassifizierung durchführen, allerdings werden noch einige Seiten falsch klassifiziert. Diese Fehlerquote kann mit kontextsensitiven Merkmalsvektoren weiter verringert werden, so dass damit die besten Ergebnisse erzielt wurden. Eher ungeeignet sind semantische Merkmale, da zum einem nicht alle Seiten HTML-Meta-Tags anbieten, aber auch so das Ergebnis zu wünschen übrig lässt. Für eine Klassifikation mit RDF und OWL waren fast keine Daten vorhanden, so dass damit keine Klassifizierung durchgeführt werden konnte..

DATUM DER PRÄSENTATION UND VERTEIDIGUNG	27.08.2010	GEBÄUDE UND RAUM DER VERTEIDIGUNG	GIII/02
FACHLICHE AUSRICHTUNG	Maschinelles Lernen		