Zürcher Hochschule
für Angewandte Wissenschaften

School of
Engineering

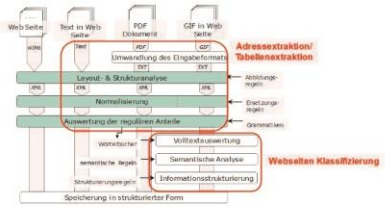
InIT Institut für angewandte
Informationstechnologie

Analysieren von semi-strukturierten Daten

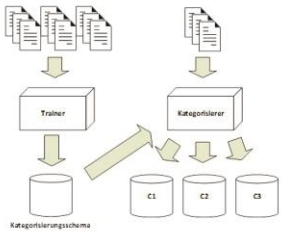
Information liegt allgemein in strukturierter, unstrukturierter oder semi-strukturierter Form vor. Dabei werden zum Beispiel Adressen oder Tabellen als strukturiert und Fliesstexte als unstrukturiert bezeichnet. Semi-strukturierte Dokumente beinhalten beides, jedoch sind die strukturierten Teile nicht explizit ausgezeichnet. Das heisst, die Metainformation fehlt, welche die Struktur beschreibt. Ziel dieser Arbeit ist es, diese Strukturen wieder aufzubauen und für die weitere Verarbeitung zugänglich zu machen. Dabei soll der Fokus auf die Erkennung und Extraktion von Adress- und Tabellenstrukturen gelegt werden. Diese Informationen mussten zur Weiterverwendung bisher mühsam von Hand aus den Dokumenten kopiert werden. Dies ist ein sehr zeitaufwendiger Prozess, welcher mit hohen Kosten verbunden ist. Um diesen Prozess zu optimieren, wurden in dieser Bachelorarbeit Methoden entwickelt, die Adressen und Tabellen erkennen und diese in einem wiederverwendbarem strukturiertem Format (XML) ablegen.
Für die Extraktion von Adressen aus HTML-Dokumenten wird die bereits bestehende Regex Markup Language (RML) angewendet, mit der die zu erkennenden Adressmuster anhand einer Grammatik definiert werden können. Ein weiterer Schwerpunkt dieser Arbeit liegt auf der Erkennung und Extraktion von Tabellen aus PDF-Dateien. Insbesondere die Tabellen in Bilanzen sind von grossem Interesse, da sie wichtige Informationen zu den Geschäftsgängen von Firmen darstellen. In Anbetracht dessen wurde ein Algorithmus entwickelt, welcher Tabellen mitsamt dem Inhalt erkennt und extrahiert. Damit die Daten aus den Tabellen weiter verarbeitet werden können, werden diese als XML-Struktur abgelegt. Schlussendlich wurde ein Webseiten-Klassifizierer implementiert, der anhand gelernter Testdaten Webseiten einer Kategorie zuteilen kann. Die Klassifizierung wurde unter anderem mit der Suchmaschine Lucene von Apache realisiert, mit der die Relevanz eines Dokumentes zu einer Kategorie berechnet werden kann.
Mit dieser Arbeit konnte aufgezeigt werden, dass es möglich ist, Adressen und Tabellen anhand ihrer Struktur aus semi-strukturierten Dokumenten zu extrahieren und für die Weiterverarbeitung aufzubereiten. Wenn nun diese Erkenntnisse mit den Möglichkeiten der Klassifikation von Dokumenten kombiniert werden, kann in Zukunft ein nützliches Information-Retrieval-System entstehen.

Diplomand
Jan Stucky

Dozent
Karl Rege



Der Information-Extraction-System kann strukturierte Daten, wie Adressen und Tabellen, aus einem semi-strukturierten Dokument extrahieren.



Der Kategorisierer ist in der Lage, aus einer Sammlung von Testdaten zu lernen. Dieses Wissen kann verwendet werden, um Webseiten zu kategorisieren.