# Extraktion der Tabellen aus XML-Dokumenten und Erkennung deren Semantik

## Exposé zur Bachelorarbeit

eingereicht von Irina Glushanok
23.04.2015

# 1 Einführung

Um eine bequeme Suche nach passender Literatur zu ermöglichen, bieten viele Bibliotheken die Möglichkeit der elektronischen Suche. Bei einer detaillierten Suche kann der Anwender festlegen, welche Schlagwörter im Text eines Dokuments und welche in seinen Metadaten, d. h. Titel, Autor(en), Verlag etc., gefunden werden sollen. Eine explizite Suche nach Tabellen und deren Analyse wurde jedoch eine lange Zeit nicht unterstützt. Dabei sind Tabellen ein fester Bestandteil von wissenschaftlichen Publikationen und dienen i. d. R. einer kompakten Darstellung wissenschaftlicher Erkenntnisse bzw. Versuchsergebnisse. In dieser Bachelorarbeit soll die Extraktion von Tabellen und die Erkennung deren Semantik implementiert werden.

# 2 Ziel

Es wird mit dem Corpus von BioMed Central [5] gearbeitet. Es beinhaltet ca. 250.000 Volltextpublikationen aus den Domänen Biologie, Medizin und Biomedizin. Alle Publikationen liegen im xml-Format vor.

Diese Bachelorarbeit umfasst die nachfolgenden Teilaufgaben:

1. Extraktion von Tabellen
2. Auslesen von tabellenbezogenen Daten (Spalten- und Zeilenüberschriften, Zelleninhalte, Tabellenüberschrift, Absatz, in dem auf die Tabelle verwiesen wird)
3. Gesamtdatensatz statistisch beschreiben
4. Clustern von Tabellen mit dem Ziel, zusammenhänge Tabellen zu identifizieren
5. Visualisierung der Cluster
6. Evaluation des Clusterings

# 3 Ansatz

Alle Artikel im o. g. Corpus unterliegen einem XML-Schema, sodass sich die ersten zwei Teilaufgaben mittels XQuery und XPath mit

einem überschaubaren Aufwand lösen lassen. Somit liegt der Schwerpunkt der Arbeit beim Punkt 3.

Die Erkennung der Semantik von Tabellen entspricht der im Data-Mining-Bereich bekannten Clustering-Aufgabe. Das Ziel ist also, die Termvektoren, die die vorliegenden Tabellen repräsentieren, in semantisch zusammenhängende Cluster so zu gruppieren, dass der Abstand zwischen den Clustern maximiert und der Abstand zwischen den einzelnen Termvektoren innerhalb jedes Clusters minimiert wird. Danach müssen für jedes Cluster Schlagwörter gefunden werden, die seine Semantik widerspiegeln. Zur Lösung der Clustering-Aufgabe haben sich einige Verfahren etabliert. Darunter der $k$-Means-Algorithmus, der in dieser Arbeit angewandt wird. Das $k$ steht hier für die Anzahl der Cluster.

Eine wichtige Voraussetzung für die Implementation des o. g. Algorithmus ist die s. g. Feature Selection. Alle zu untersuchenden Objekte, in unserem Fall Tabellen, werden durch die Termvektoren gleicher Länge repräsentiert. Die Werte der einzelnen Komponenten dieser Vektoren indizieren, mit welcher Gewichtung ein Term in den tabellenbezogenen Daten vorhanden ist. Es ist wichtig, die Dimension des Vektorraums zu reduzieren und sich auf die Terme einzuschränken, die die Unterschiede zwischen den Objekten erkennbar machen. Mit diesen Maßnahmen erreicht man eine Reduktion der Anzahl der Rechenoperationen und eine Qualitätserhöhung des Clusterings.

Die nachfolgenden Techniken könnten zum Einsatz kommen, um die Feature Selection zu unterstützen:

1. Stop-Wörter entfernen
2. numerische Werte entfernen, da diese nicht zur Semantikerkennung beitragen
3. die Terme aus den Absätzen, die auf die Tabellen verweisen, nutzen
4. Das Feature Subset kann durch die entsprechenden Oberbegriffe von vorkommenden Instanzbezeichnungen angereichert werden. Dies kann mit Hilfe eines Thesaurus erfolgen, falls frei verfügbar.

Ob die Instanznamen selbst aus dem Feature Subset entfernt werden sollten, muss noch untersucht werden.

5. Reduktion der Dimensionen der Termvektoren durch Hauptkomponentenanalyse (engl. Principal Component Analysis oder kurz PCA)

Ferner muss für die Anwendung des $k$-Means-Algorithmus ein geeignetes $k$ gewählt werden, welches als Parameter an den Algorithmus übergeben wird. Ein Verfahren für die Bestimmung von $k$ wird in [3] beschrieben. Man initialisiert das $k$ mit einem kleinen Wert, z. B. mit 1. In jedem Iterationsschritt wird der $k$-Means-Algorithmus ausgeführt. Dann wird testet, ob die Punkte in jedem erzeugten Cluster normalverteilt sind. Im positiven Fall wird die Schleife abgebrochen, im negativen Falls wird das $k$ erhöht. Die Prüfung der Punkte auf Normalverteilung erfolgt mittels Anderson-Darling-Test.

Für die Clustervisualisierung scheinen die Parallelkoordinaten [4] ein geeignetes Mittel zu sein. Sie bringen den Vorteil, dass man durch die Beibehaltung der Dimensionen der bereits durch Preprocessing aufbereiteten Termvektoren keine für den Betrachter wichtige Information verliert. Für jeden Termvektor wird eine Art Zickzack-Kurve erzeugt, die die Koordinatenpunkte der einzelnen Vektorkomponenten verbindet. Die Cluster sollten als Kurvenbündel zu erkennen sein.

## 4 Related Work

Es gibt bereits einige nennenswerte Arbeiten auf dem Gebiet der Tabellensuche. Die meisten konzentrieren sich auf die Suche von Tabellen in semistrukturierten Dokumenten. So bietet z. B. Google die Suche nach Tabellen in den Webseiten [6]. Das Projekt ist aber noch im Entwicklungsstadium.

Liu et al. [1] haben ein Tool zur Suche von Tabellen in PDF-Dokumenten entwickelt. Die Tabellenerkennung erfolgt regelbasiert. Das Ranking von Tabellen wurde klassisch über das Vektorraum-Modell realisiert, jedoch wurden die üblichen Term-Gewichtungen TF und IDF an die Aufgabenstellung angepasst und wurden zu s. g. TTF (Table Term Frequency) und ITTF (Inverse Table Term Frequency).

Einen anderen Schwerpunkt hat die Arbeit von S. Kim et al. [2]. Hier liegt die Problematik der Klassifikation von Tabellen im Fokus. Jede Tabelle wird einer der vordefinierten Klassen zugeordnet, wie z. B. 'Statistics' oder 'Example'. Auch hier hat man mit den in [1] erwähnten speziellen Termgewichtungen gearbeitet.

## References


[1] Ying Liu, Kun Bai, Prasenjit Mitra, C. Lee Giles: TableSeer: Automatic Table Metadata Extraction and Searching in Digital Libraries

[2] Seongchan Kim, Keejun Han, Soon Young Kim, Ying Liu: Scientific Table Type Classification in Digital Library

[3] Greg Hamerly, Charles Elkan: Learning the k in k-means

[4] Edward J. Wegman: Hyperdimensional Data Analysis Using Parallel Coordinates. Journal of the American Statistical Association, Vol. 85, No. 411 (Sep., 1990), 664-675.

[5] BioMed Central `http://www.biomedcentral.com/`

[6] Google Table Search `https://research.google.com/tables`