TEXT MINING BEI GERICHTSENTSCHEIDUNGEN

Matthias Prinz

Wissenschaftlicher Mitarbeiter und Rechtsanwalt, Technische Universität Darmstadt, Lehrstuhl Prof. Dr. Marly
Hochschulstraße 1, 64298 Darmstadt. DE
mail@prinz.law; www.zivilrecht.wi.tu-darmstadt.de

Schlagworte: ***Text Mining, Rechtsausführungen, Rechtsnormverweise, Rechtsvisualisierung***

Abstract: *Datengestützte Rechtsinformatik ist auf die Analyse von Rechtstexten und die automatische Erfassung daraus nutzbarer Daten angewiesen. Vorgestellt wird eine Analysesoftware, die Gerichtsentscheidungen aus gängigen Dateiformaten einliest und auswertet. Es werden solche Rechtsausführungen des Gerichts auf Satzebene gesammelt, die vom zugrundeliegenden Tatbestand losgelöst sind. Strategien zum Erkennen dieser Ausführungen werden vorgestellt. Ebenso wird exemplarisch aufgezeigt, wie diese Rechtsausführungen Juristen entweder systematisch, kontextabhängig oder graphisch zugänglich gemacht werden können.*

1. Aufbau der Analysesoftware

Die quantitative Analyse von Gerichtsurteilen ist aufgrund der Textmenge nur computergestützt möglich. Daher besteht die Notwendigkeit zur Entwicklung von geeigneter Software. So sollen Gerichtsentscheidungen aus unterschiedlichen Dateiformaten in ein einheitlich zu bearbeitendes Format übertragen werden können.

Als Basis sollten die offensichtlich verfügbaren Informationen wie Entscheidungsdatum, Aktenzeichen, Gericht, Gerichtstyp, Entscheidungstyp, Gesetzesreferenzen und Referenzen auf andere Entscheidungen erkannt und strukturiert gespeichert werden. Um auf Satzebene rechtliche Maßstäbe erkennen zu können, ist zudem eine Segmentierung des Entscheidungstextes in einzelne Sätze notwendig.

Umgesetzt wird dies mit einer entwickelten Software, welche auch Open-Source Software als Programmbibliotheken einsetzt und in der Programmiersprache Java implementiert ist.

Hilfreich sind dazu insbesondere folgende Software-Drittbibliotheken:

- «Apache Tika»[1], um aus den verschiedenen Dateiformaten wie PDF mit hinterlegter Schrift, einfachen Textdateien, oder formatierten Texten (rich text format) einen Text als Zeichenkette (String) zu erhalten.
- Für die Satzgrenzenerkennung kommt im ersten Schritt die Programmbibliothek «Apache Open NLP» zum Einsatz, die unter Verwendung von machine learning die Satzgrenzen zu erkennen versucht. Problematisch sind die vielen Abkürzungen in juristischen Texten, die zu falscher Satzgrenzenerkennung führten. Dieses Problem konnte weitgehend durch eine Abkürzungsliste behoben werden.

Für das Datenmodell wurden einfache Java Objekte und eine dokumentenbasierte Struktur gewählt. Das Wurzelelement jeder Entscheidung ist das Objekt *decision*, welches den Text, die Metadaten und weitere Unterobjekte wie eine Liste von Zitierverweisen oder Gesetzesverweisen enthält. Nach dem Einlesen und Extrahieren der Informationen lassen sich die Daten entweder in Datenbanken oder zum Datenaustausch in einem JSON[2]-Format weiternutzen. Dieses JSON Format kann einfach aus den bestehenden Java Objekten generiert werden. Hierzu eignet sich die Programmbibliothek «Gson» von Google[3].

---

1 https://tika.apache.org/ (alle Websites zuletzt abgerufen am 24. Januar 2018).

2 JSON steht für JavaScript Object Notation und wird oft beim Austausch von Daten zwischen Anwendungen genutzt.

3 https://github.com/google/gson.

Nachdem der Entscheidungstext nach Einlesen der Datei als Zeichenkette vorliegt, kann die Software den Text von unerwünschten mehrfachen Zeilenumbrüchen, Seitennummerierungen, offensichtlich fehlerhaften Zeichen und unnötigen Worttrennzeichen bereinigen.

Danach werden die oben genannten Metainformationen wie unter anderem das zu erkennende Gericht, Aktenzeichen, Entscheidungsdatum und Entscheidungstyp hinzugefügt. Die Extraktion dieser Metadaten erfolgt mittels Beschreibung der Daten über sog. Reguläre Ausdrücke. Dies hat sich beim Entwickeln der Software als recht zuverlässig erwiesen. Problematisch erscheint, dass das Nachvollziehen, Testen und Validieren solcher Regulärer Ausdrücke schwierig ist.

Teilweise stößt aber die Erkennung mit Regulären Ausdrücken an praktische Grenzen. Beispielsweise erscheint es einfacher, zur Erkennung von Gesetzesreferenzen in den Entscheidungen einen eigenen Algorithmus zu implementieren, anstatt einen sehr komplexen Regulären Ausdruck zu entwickeln. Der Algorithmus zerlegt bei Vorkommen der Zeichen «§» oder «Art.» die nachfolgenden Zeichen und erkennt Paragraph, Absatz, Satz, Nummer, Halbsatz, Buchstabe und Variante sowie das Gesetzbuch Schritt für Schritt. Damit können auch Verweisketten wie z.B. «§§ 1004 Abs. 1 S. 1 analog, 823 Abs. 1 BGB» erkannt werden. Die Gesetzbucherkennung wird mit einer Tabelle von Gesetzesbezeichnungen und Alternativbezeichnungen mittels einer hashbasierten Key-Value Datenstruktur umgesetzt. Dies bietet den Vorteil, dass trotz einigen tausenden Einträgen die Laufzeit des Algorithmus zur Erkennung des Gesetzbuches ungeachtet der Anzahl der Einträge im Regelfall konstant bleibt.

Das dargestellte Vorgehen ermöglicht dann die weitergehende Analyse der Entscheidungen. So können beispielsweise Kookkurrenzen bei Rechtsnormverweisen[4] zur Erstellung eines Gesetzesgraphen[5] erfasst werden. Ferner können solche Sätze extrahiert werden, die abstrakt von dem Streit der zugrundeliegenden Entscheidung sind. Meist handelt es sich bei solchen abstrakten Sätzen um rechtliche Maßstäbe, welche die Anwendung des Rechts konkretisieren. Diese abstrakten Sätze sind oft das, was bei der Recherche von Gerichtsentscheidungen von Interesse ist. Insofern verspricht es großen Nutzen, diese Sätze zu extrahieren und stukturiert, beispielsweise in einer Datenbank, zu speichern. Andere Arbeiten haben sich im Schwerpunkt mit dem Extrahieren von Definitionen aus Entscheidungs- oder Gesetzestexten beschäftigt.[6] Die Definitionen von Begriffen sind aber enger gefasst als allgemeine Ausführungen zum Recht oder zur Gesetzesanwendung.

Die vorgestellte Software arbeitet insgesamt relativ schnell. Auf einem aktuellen Desktop-Computer mit vier Rechenkernen können etwa 20–40 Urteile pro Sekunde eingelesen, analysiert und abgespeichert werden.

2. Verfahren zur Gewinnung rechtlicher Maßstäbe

Sätze, die rechtliche Maßstäbe enthalten, sind zumindest in Teilen vom zugrundeliegenden Streit der Parteien losgelöst. Die erkennenden Richter belegen Ihre Aussagen oft mit Zitaten aus anderen Entscheidungen, mit Kommentaren oder Zeitschriftenaufsätzen. Eingefasst sind die Zitate üblicherweise und fast ausschließlich mit einer Umklammerung. Findet sich nun ein Satz, der eine Umklammerung von Text enthält, prüft die Software, ob eine bekannte Zeitschrift, eine Rechtsnorm, eine andere Entscheidung oder ein Kommentar darin enthalten sind. Ist dies der Fall, kann angenommen werden, dass eine Rechtsansicht belegt wird. Ist der Satz nicht im

[4] Landthaler/Waltl/Matthes, Differentiation and Empirical Analysis of Reference Types in Legal Documents, in: Jusletter IT Flash 17. August 2017 beschreiben zutreffend, dass auch sog. stillschweigende Verweise (tacit references) wie Analogien im Gesetzestext durch die Analyse von Rechtsnormreferenzen in Gerichtsentscheidungen oder Sekundärliteratur erschlossen werden.

[5] Bei Kookkurrenzen von Gesetzesverweisen innerhalb einer Entscheidung wird pauschal eine inhaltliche Beziehung unterstellt und eine Verbindung zwischen den Normen angelegt. Dies mag im Einzelfall nicht zutreffend sein. Erhöht sich das «Kantengewicht» mit zunehmender Anzahl der Kookkurrenzen von zwei Rechtsnormen und filtert die Verbindungen mit dem höchsten Verbindungsgewicht, so erscheinen aus dem Studium bekannte Verweisstrukturen, die sich aber selbst im Gesetzestext nicht wiederfinden (siehe Abb. 2).

[6] Waltl/Landthaler/Scepankova/Matthes/Geiger/Stocker/Schneider 2017; Walter 2009.

Konjunktiv gefasst, erhöht dies die Wahrscheinlichkeit, dass es sich um eine Rechtsausführung des Gerichts handelt. Ebenfalls sind in den allermeisten Fällen die abstrakten Rechtsausführungen im Präsens verfasst.

Nicht alle Rechtsausführungen enthalten Zitate. Daher lässt sich das Verfahren vermutlich dadurch weiter verbessern, indem unter Einsatz von Text-Classifier weitere abstrakte Sätze identifiziert werden. Solche Verfahren des maschinellen Lernens benötigten üblicherweise Beispiele, in der Form von Text und Zuordnung (supervised machine leanring). Sätze mit Rechtsausführungen ohne Zitate dürften strukturähnlich zu denen mit einem Zitat sein, wenn der Text mit Zitat in der Klammer entfernt wird. Unter dieser Prämisse kann man die bereits erfassten Sätze als Trainingsdaten für Text-Classifier nutzen. Zusätzlich bieten sich noch die amtlichen Leitsätze der Bundesgerichte als Trainingsdaten an.[7] Bei englischsprachigen Gerichtsentscheidungen lassen sich rechtliche Prinzipien von Tatsachen mithilfe solcher Verfahren abgrenzen.[8] Auch (deutsche) Rechtsnormen lassen sich mit hoher Wahrscheinlichkeit in verschiedene Arten (wie Anspruch oder Einrede) richtig kategorisieren.[9]

Die eingesetzte Software ist bislang aber noch nicht so weit entwickelt, dass sie die gewünschte Kategorisierung mit Text-Classifiern über Experimente hinaus umzusetzen kann, sondern beschränkt sich zum jetzigen Zeitpunkt auf das Herausfiltern von Sätze mit Zitaten.

Um den Kontext des extrahierten Satzes zu bewahren, werden vorhergehende Sätze und nachfolgende Sätze als Textfenster sowie eine Referenz auf die zugrundeliegende Entscheidung mitgespeichert.

Das aufgezeigte Verfahren grenzt sich von anderen Vorschlägen ab, die auf das Extrahieren von Definitionen für einzelne Tatbestandsmerkmale abzielen.[10] In die gleiche Richtung wie das hier vorgestellte Verfahren geht die Arbeit von Carlson mit der LawProp[11] Software, die jedoch auf die originale Belegstelle über Verweise und Textähnlichkeit in anderen Entscheidungen die Textstelle mit «legal propositions» in der zitierten Entscheidung finden kann. Dieses Vorgehen scheint ebenfalls auf deutsche Gerichtsentscheidungen übertragbar zu sein.

Dem hier dargestellten Vorgehen ist allerdings einschränkend anzumerken, dass Sätze sowohl streitbezogene Teile als auch abstrakte Rechtsausführungen beinhalten können. Es besteht daher Bedarf, die Erkennung von abstrakten Rechtsausführungen auch auf Satzteile anwenden zu können. Darüber hinaus ist das Problem des Erkennens von Rechtsausführungen über Satzgrenzen hinweg noch offen.

Schließlich ist zu prüfen, ob das Übernehmen von Rechtsausführungen einen nicht gerechtfertigten Eingriff in die Urheberrechte Dritter darstellt, insbesondere wenn Gerichtsentscheidungen private (Literatur-)Werke wie Kommentare oder Aufsätze in Gerichtsentscheidungen zitieren.

Derjenige, der per Text Mining gewonnene Rechtsausführungen nutzt, könnte sich möglicherweise auf das Zitatrecht nach § 51 UrhG berufen. Die Regelbeispiele des § 51 UrhG für zulässige Zitate sehen allerdings sämtlich die Verwertung in einem eigenen, selbständigen Werk vor. Zudem ist § 51 UrhG als Schrankenregelung grundsätzlich eng auszulegen. Durch das automatische Sammeln der Rechtsausführungen durch den Computer wird kein eigenes urheberrechtliches Werk geschaffen, da es an der erforderlichen persönlichen geistigen Schöpfung nach § 2 Abs. 2 UrhG fehlt. Insofern kann sich derjenige, der die Rechtsausführungen per Text Mining nutzen möchte, nicht auf die Schranke des Zitats nach § 51 UrhG berufen.

Urteile sind allerdings nach § 5 Abs. 1 UrhG als amtliche Entscheidungen gemeinfrei. Völlig unproblematisch ist es daher, wenn Rechtsausführungen von anderen Gerichten zitiert werden und durch die Software gesammelt werden. Gleiches gilt für amtliche Leitsätze, z.B. von der Dokumentationsstelle der Gerichte.[12]

---

7 Diese lassen sich aus den XML-Entscheidungsdateien von Rechtsprechung-im-Internet.de automatisch extrahieren.

8 Shulayeva/Siddharthan/Wyner 2017.

9 Waltl/Muhr/Glaser/Bonczek/Scepankova/Matthes 2017.

10 Walter 2009.

11 Carlson 2015.

12 Vgl. VGH Mannheim, Urteil vom 7. Mai 2013 – 10 S 281/12 = GRUR 2013, 821.

Enthalten Gerichtsentscheidungen allerdings Zitate privater Werke, sind die zitierten Texte dadurch nicht (völlig) gemeinfrei. So bleiben diese im Rahmen der allgemeine Schranken im Original weiter geschützt. Dritten darf das Werk ohne Zustimmung des Urhebers bzw. Rechteinhabers nur im Kontext des amtlichen Werkes, ansonsten aber nur im Rahmen der sonstigen Schranken des Urheberrechts verwertet werden.[13] Da die extrahierten Zitate den Kern der Entscheidungsbegründung betreffen, erfasst dies den Kerngehhalt des § 5 UrhG, nämlich das öffentliche Interesse der Bevölkerung an allgemeiner Kenntnisnahme. Daher ist es erforderlich, dass die Zitatstellen, wie durch das Gericht zitiert, gemeinfrei werden, solange der Bezug zum amtlichen Werk bestehen bleibt. Da die allgemeine Kenntnisnahme der Begründungen der Entscheidung sichergestellt sein soll, muss es daher ausreichen, wenn ein Link von der Rechtsausführung zu dem entsprechenden Urteil im Volltext führt.

Bei der Nutzung der durch Text Mining gewonnenen rechtlichen Maßstäbe werden daher keine Urheberrechte Dritter verletzt, solange der Bezug zum Urteil erkennbar bleibt.

3. Darstellung der rechtlichen Maßstäbe

3.1. Als Annotation zum Gesetzestext

Enthalten die rechtlichen Maßstäbe selbst eine Gesetzesreferenz können diese – wie in einem Gesetzeskommentar – zur Erläuterung des Gesetzestextes herangezogen werden. Nachteilig ist die fehlende inhaltliche Aufbereitung der Fundstellen untereinander. Diese Rechtsinformationen dürften daher von geringerer Qualität sein. Es sind jedoch die hohen Kosten für den Zugang zu juristischer Sekundärliteratur zu bedenken. Daher bietet dieses Vorgehen dann einen Mehrwert, wenn kein Zugang zu juristischer Sekundärliteratur besteht.

3.2. Per Volltextsuche

Die gewonnen rechtlichen Maßstäbe können entweder über eine Volltextdatenbank[14] durchsucht oder kontextabhängig beim Schreibprozess genutzt werden. Dies könnte hilfreich sein, wenn lediglich eine bekannte Rechtsansicht nochmals rekapituliert werden soll oder ein Beleg für die Rechtsansicht benötigt wird.

Die nachfolgende Grafik zeigt eine entsprechende Word Erweiterung, die beim Markieren innerhalb des Dokumententextes einen Suchvorgangs von Sätzen mit rechtlichen Maßstäben im rechten Fenster auslöst. Die rechtlichen Maßstäbe können per Mausklick samt Zitat auf die Entscheidung in das Dokument des Nutzers übernommen werden.

[13] Nordemann, in Fromm/Nordemann 2014, zu § 5, Rn. 13.

[14] Die Daten werden in die Datenbanken vom Typ MongoDB und Elasticsearch eingespielt.

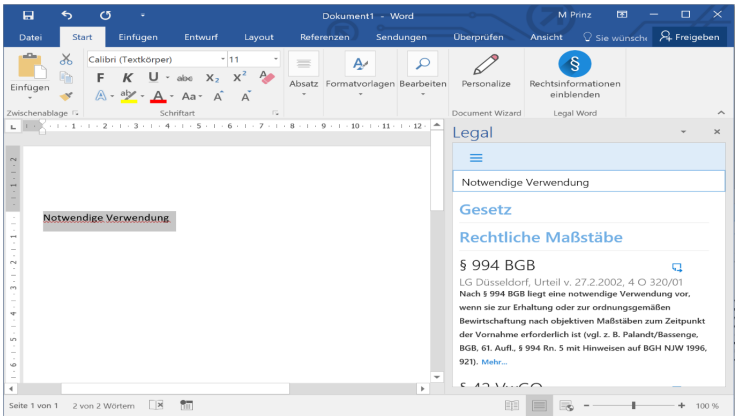

Abbildung 1: Word Add-In «Legal» mit rechtlichen Maßstäben.

3.3. Per Annotation von Rechtsnormen untereinander

Bereitet man die Rechtsnormen grafisch als Netzwerk auf, so können auf Verbindungslinien zwischen Gesetzen solche Rechtsinformationen eingefügt werden, die entweder mindestens zwei Rechtsnormreferenzen enthalten oder die Rechtsnormen davor und danach im Entscheidungstext nennen. Dies bietet den Vorteil, dass sich die Verbindungen qualitativ beschreiben lassen. Beim Einarbeiten in neue Rechtsgebiete oder für eine Übersicht könnte ein solcher Gesetzesgraph einen zusätzlichen Rechercheeinstieg ermöglichen, insbesondere wenn die Rechtsinformationen wie Gesetze und rechtliche Maßstäbe beim Auswählen der Gesetze oder der Verbindung zwischen zwei Gesetzen angezeigt werden.

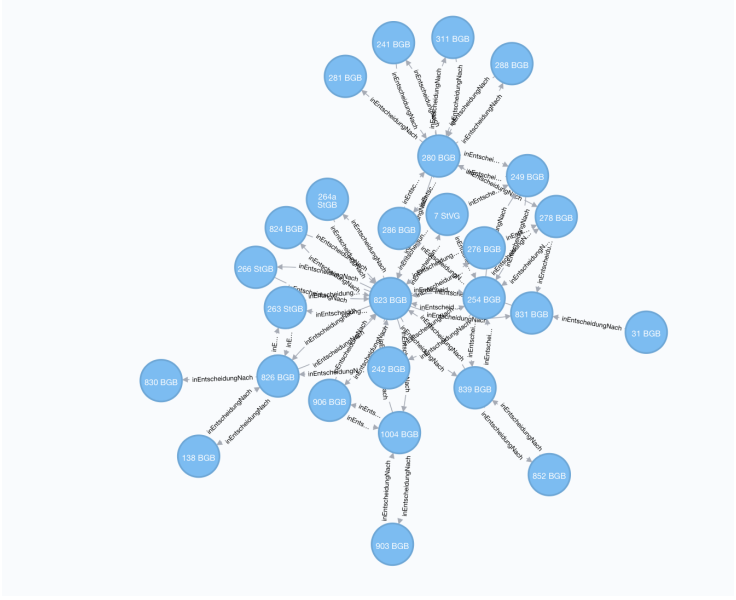

Abbildung 2: Gesetzesnetzwerk zu § 823 BGB. Verbindungen aus Kookkurrenzen von Gesetzesverweisen, gefiltert nach BGB, StGB und StVG und sortiert nach Verbindungsstärke.

4. Verfügbare Gerichtsentscheidungen

Es wurden verschiedene Korpora von Gerichtsentscheidungen mit insgesamt knapp 300'000 Entscheidungen zusammengetragen. Die einzelnen Korpora stammen aus folgenden Quellen: Bundesgerichtshof (98'000 Entscheidungen), Bundesverfassungsgericht, Bundesarbeitsgericht, Bundesverwaltungsgericht, Bundessozialgericht, Bundespatentgericht, der online Rechtsprechungsdatenbank Bayern, der online Rechtsprechungsdatenbank NRW sowie Entscheidungen des Europäischen Gerichtshofs auf curia.europa.eu.

Vor allem ältere Entscheidungen von Bundesgerichten lagen als Quelldatei nur als Grafik aus Scans in PDF-Dateien vor. Solche Dateien wurden mit OCR-Software zunächst in PDF-Dateien mit hinterlegter Textebene umgewandelt. Bei älteren Entscheidungen, insbesondere vor 1980, häufen sich die Erkennungsfehler aufgrund qualitativ schlechter Schreibmaschinenschrift.

Ab dem Jahr 2010 stehen die meisten Entscheidungen der Bundesgerichte auf Rechtsprechung-im-Internet.de die in einem strukturierterem XML-Format zum Download bereit.

Die vorgestellte Software kann alle Entscheidungen einlesen, verarbeiten und strukturiert speichern.

5. Literatur

Carlson, James, LawProp, Using Quotations to Identify Legal Propositions in Judicial Opinions, Stanford University: CODEX: The Stanford Center for Legal Informatics, October 26, 2015.

Fromm, Axel/Nordemann, Jan Bernd (Hrsg.) Urheberrecht, 11. Auflage, 2013.

Landthaler, Jörg/Waltl, Bernhard/Matthes, Florian, Differentiation and Empirical Analysis of Reference Types in Legal Documents, in: Jusletter IT Flash 17. August 2017.

Shulayeva, Olga/Siddharthan, Advaith/Wyner, Adam, Recognizing cited facts and principles in legal judgements, Artificial Intelligence and Law, vol. 25, 2017.

Walter, Stephan, Definition extraction from court decisions using computational linguistics technology, Formal Linguistics and Law, vol. 212, 2009.

Waltl, Bernhard/Landthaler, Jörg/Scepankova, Elena/Matthes/Geiger/Stocker/Schneider, Automated

Waltl, Bernhard/Muhr, Johannes/Glaser, Ingo/Bonczek, Georg/Scepankova, Elena/Matthes, Florian, Classifying Legal Norms with Active Machine Learning, Jurix: International Conference on Legal Knowledge and Information Systems, Luxembourg, Luxembourg, 2017, S. 11–21.