



Automatische Kategorisierung Juristischer Dokumente

Anton Geist
Arbeitsgruppe Rechtsinformatik
Wiener Zentrum für Rechtsinformatik

Übersicht

- Die Definition & Das Problem
- Die Beiden Beispiele
- LexisNexis: *Term-Based Topic Identification System (TTI)*
- Westlaw: *Classification and Recommendation Engine (CaRE)*
- Westlaw Beispielanwendung: *ResultsPlus*
- Meine Thesen & Unsere Diskussion

Die Definition & Das Problem

- Dokumentenkategorisierung =
- Automatisches Erkennen und Hinzufügen von zusätzlichen Dokumenteninformationen, zur Kategorisierung, beziehungsweise zur Verbesserung der Suche allgemein
- Kategorisierung = (eine Art von) Indexierung
- Manuelle Kategorisierung
 - teuer
 - fehlerhaft

Die Beiden Beispiele

- Westlaw und LexisNexis dominieren als Vorreiter der elektronischen juristischen Recherche (*computer-assisted legal research*) nach wie vor den US-amerikanischen, und teilweise auch den internationalen, Markt (“Wexis”).
- **LexisNexis**
5+ Milliarden Dokumente online (Stand 2009)
Datenwachstum 16 Dokumente / Sekunde (Stand 2001)
- **Westlaw**
500 000 zu kategorisierende Rechtssätze pro Jahr
(Stand 2007)

LexisNexis: Term-Based Topic Identification System (TTI)

- *Controlled Vocabulary Terms (CVTs)* werden durch Gewichtung der Dokumentinhalte identifiziert
 - Die Gewichtung erfolgt eher auf Konzept-, als auf Termebene
- Das System arbeitet praktisch ohne Stoppwörter.
- *Topic Scores* beschreiben den relativen Umfang eines Themas im Dokument.
- Der Aufbau einer Klassifizierungs-Regelgruppe dauert zwischen 5 Minuten und 8 Stunden.
- Die erreichten Precision (Treffericherheit) und Recall (Vollständigkeit) Werte liegen für die meisten Themen zwischen 90 und 95%.

Westlaw: Classification and Recommendation Engine (CaRE)

- West Enzyklopädien, Kommentare sowie das West *Key Number System* enthalten – in unterschiedlichen Kategoriensystemen – Verweise auf Rechtssätze.
- Die rasche, richtige und flexible Kategorisierung von Rechtssätzen ist daher das Um und Auf.
- Die CaRE Classification and Recommendation Engine nutzt mehrere Klassifikatorenmodule:
 - Support Vector Machine (SVM)
 - Bayes-Klassifikator
 - Nächste-Nachbarn-Klassifikation
- Genutzt werden jedenfalls bereits vorhandene, richtige Kategorisierungen zum Aufstellen von Regeln (Lernen) und dann zum Kategorisieren von neuen Dokumenten.

Westlaw Beispielanwendung: ResultsPlus

- 2002 entwickeltes Vorschlagssystem, das auf CaRE aufsetzt
- bei Judikatorsuchen wird einschlägige Literatur (Zeitschriften, Kommentare) dynamisch vorgeschlagen
- Entwicklungszeit: 6 Monate
- Entwicklungskosten bis zum Einsatz: 1 Million USD
- Einnahmen im ersten Jahr im Einsatz: 2 Millionen USD
- Zugerechnete Mehreinnahmen (Stand 2008):
40 Millionen USD

Meine Thesen & Unsere Diskussion

- Bereits vorhandene, überprüfte Kategorisierungen sollten für ein automatisches System genutzt werden.
- Für die Ermittlung von Kategorisierungen ist die Nutzung juristischer Konzeptlisten unbedingt erforderlich.
- In jedem System muss der Recall/Precision Trade-Off beachtet werden.
- Für jede Rechtsdatenbank mit mehreren Suchkategorien wäre eine Funktion wie ResultsPlus eine Success Story.