Zusammenfassung

der Dissertation "On Information Need and Categorizing Search" von S. Meyer zu Eißen

Das Fachgebiet Information Retrieval (IR) beschäftigt sich mit der computergestützten Erfüllung von Informationsbedürfnissen. Typische IR-Systeme nehmen Anfragen eines Benutzers entgegen und versuchen, diese auf Basis von großen Dokumentmengen zu beantworten. Techniken zur Befriedigung von Informationsbedürfnissen sind mittlerweile allgegenwärtig, sei es in Form von Suchmaschinen daheim oder bei der Arbeit, auf mobilen Geräten oder auf Workstations, oder auch als Retrieval-Komponenten in File-Systemen, Dokument-Ablagen, etc. Zu den Gründen dieser starken Verbreitung von IR-Systemen zählen unter anderem das wachsende Informations-bedürfnis, die Vielfältigkeit der Retrieval-Aufgaben und der gewünschte Grad an Personalisierung. In der Dissertation werden neue Konzepte und Algorithmen entwickelt, die den gesamten IR-Prozess betreffen; insbesondere aber die automatische Kategorisierung von Dokumentmengen gemäß Thema und Genre.

Dokument-Repräsentation. Traditionelle Methoden zur Messung von Dokument-ähnlichkeiten vernachlässigen die Wortreihenfolge-Information. Das vorgeschlagene Suffix-Baum-Dokumentmodell ist in der Lage, sowohl Termübereinstimmungen als auch Reihen-folgeübereinstimmungen zu quantifizieren. Experimente zeigen, dass das Suffix-Baum-Modell die Qualität von unüberwachten Dokument-Kategorisierungen stark verbessern kann.

Cluster-Validität. In Szenarien, in denen unüberwacht Dokument-Kategorien gebildet werden, sind üblicherweise die Parameter der zugrundeliegenden Clustering-Algorithmen unbekannt und müssen (wiederholt) geschätzt werden. Wir schlagen das Cluster-Validitäts-Maß $\overline{\rho}$ vor, das in der Lage ist, Clusterings gemäß ihrer Qualität gemessen an menschlichen Vorstellungen zu beurteilen.

Topic-Identifikation. Nachdem eine unüberwachte Kategorisierung einer Dokumentmenge ermittelt ist, soll diese einem Benutzer z.B. zur Navigation präsentiert werden. Das bedeutet, dass die Kategorien mit charakteristischen Termen beschriftet werden müssen. Wir definieren formale Anforderungen an solche Kategoriebezeichner und schlagen den Algorithmus WCC zur Berechnung solcher Beschriftungen vor.

Genre-Kategorisierung. In der aktuellen IR-Forschung bezeichnet "Kategorisierung" oft eine Ordnung bezüglich *Thema*. In der Arbeit wird gezeigt, dass das *Genre* eines Dokuments ein sehr nützliches Kategorisierungskriterium bei der Suche in großen Dokumentmengen sein kann. Insbesondere wird ein Genre-bezogenes Kategorie-System für Web-Dokumente vorgeschlagen sowie ein Dokumentmodell zur Klassifikation von Web-Dokumenten gemäß dieses Systems. Machbarkeitsanalysen zeigen, dass Genre-Kategorisierung auch in großen, heterogenen Dokumentkollektionen durchführbar ist.

Softwaretechnik. Zur Operationalisierung von IR-Prozessen wird ein Model Driven Architecture (MDA)-Ansatz mit dem Namen TIRA vorgeschlagen, der im Gegensatz zum vorherrschenden Library-basierten Ansatz klar die IR-Prozess-Spezifikation von der Operationalisierung trennt. TIRA ermöglicht es, IR-Prozesse auf persönliche Bedürfnisse zuzuschneiden.

Die vorgeschlagenen Konzepte und Algorithmen wurden u.a. in der Meta-Suchmaschine AIsearch umgesetzt, die bei dem EASA-Wettbewerb (European Academic Software Award) 2004 ausgezeichnet wurde.