

Buchbesprechung, erschienen in: nfd Information – Wissenschaft und Praxis 2000, vol. 51 (4). S. 247 f.

**Ricardo Baeza-Yates; Berthier Ribeiro-Neto (Hrsg.):
"Modern Information Retrieval"**

Addison-Wesley und ACM Press 1999, 513 Seiten (brosch.)

Mit diesem Band wollen die Herausgeber und die weiteren Autoren ein umfassendes Lehrbuch für Information Retrieval (IR) vorlegen. Die Herausgeber lieferten gleichzeitig die meisten Beiträge, einzelne Kapitel stammen von Experten für die jeweiligen Themen. Die Autoren sind im Inhaltsverzeichnis nicht vermerkt, in diesem Text werden sie in Klammern aufgeführt, soweit es sich nicht um die Herausgeber handelt.

Der Fokus des Buchs liegt auf der Sicht der Informatik: „This means that the focus of the book is on computer algorithms and techniques used in information retrieval systems. A rather distinct viewpoint is taken by librarians and information science researchers, who adopt a human-centered interpretation of the IR problem“ (S. 3). Diese isolierte Sichtweise grenzt aber wichtige Aspekte aus. Der Benutzer spielt im IR die zentrale Rolle und das Ziel ist die Erfüllung von Informationsbedürfnissen. Dieser Grundsatz hat sehr wohl auch Auswirkungen auf Algorithmen und Techniken. So führten z.B. die Schwierigkeiten der Benutzer mit Booleschen IR Systemen zum Durchbruch von Ranking Systemen im Internet. Der Blickwinkel der Informationswissenschaft kann also nicht einfach ausgeblendet werden, ohne dass Fundamentales verloren geht.

Es ist schade, dass benutzerorientierte Ansätze in den 500 Seiten nicht stärker vertreten sind, denn ansonsten könnte der Aufbau des Buchs teilweise für den Aufbau eines Seminars dienen. Auch die im Titel versprochene Modernität wird z.B. durch die Behandlung von TREC (wichtige Evaluierungsinitiative, cf. trec.nist.org), Multimedia IR, Digitale Bibliotheken und IR im Internet eingelöst. Damit haben auch Studierende, die den Band als Einführung nutzen, sofort einen Bezug zwischen IR und modernen Themen.

Nach einem einführenden Überblickskapitel folgt ein Kapitel über Modelle, die in klassische Modelle (Boolesches, Vektorraum und Probabilistisches Modell) und alternative Modelle aufgeteilt sind. Das Boolesche Modell ist aber inzwischen nicht mehr das dominante Modell in kommerziellen Systemen wie die Autoren behaupten (S. 26). Viele grosse Retrievalsystem-Hersteller (z.B. Fulcrum, Verity) vertreiben inzwischen Ranking-Systeme, ebenso bieten die meisten Hersteller von DBMS als Zusatzmodule Text-Retrievalsysteme mit Ranking-Funktionalität an (z.B. IBM, Informix, Oracle, Sybase). In der Beschreibung des Vektorraum-Modells wünscht man sich mehr alternative Gewichtungformeln. Die alternativen Modelle sind wiederum unterteilt in alternative mengentheoretische Modelle (Fuzzy-Set, Extended Boolean), algebraische Modelle (Generalized Vector Space, Latent Semantic Indexing, Neuronale Netze) und probabilistische Modelle (Bayes'sche Netze). Von den neuronalen Netzwerk-Modellen wird nur das Spreading-Activation-Modell aufgeführt, das konzeptuell dem Vektorraum-Modell sehr nahe steht und nur über geringe Lernfähigkeit verfügen. Kohonen-Netzwerke tauchen erst unter den Benutzungsoberflächen auf, obwohl sie gut in den abschliessenden Abschnitt über Browsing passen würden. Der scheinbar radikale Schwenk von Suche zu Browsen ist durchaus gerechtfertigt und die Autoren leiten ihn aus unterschiedlichen Benutzerabsichten ab. Hier zeigt sich an einem Detail, wie grundlegend die benutzerzentrierte Sichtweise auf Informationsprozesse ist. Insgesamt ist das Modell-Kapitel gut aufgebaut und umfasst mit Ausnahme der linguistisch orientierten IR-Modelle alle wichtigen Ansätze.

Bereits das dritte Kapitel behandelt Evaluierung, die damit einen angemessenen Platz einnimmt. Allerdings fällt es etwas knapp aus und die mangelnde Benutzerorientierung zeigt sich hier besonders deutlich. Die üblichen Maße recall und precision und ihre Anwendung werden vorgestellt. Die Berücksichtigung unterschiedlicher Benutzerstandpunkte fällt zu knapp aus. Der zweite Teil des Kapitels stellt verschiedene Testkollektionen vor, wobei der TREC-Konferenz zu Recht der größte Raum gewidmet wird. Man wünscht sich allerdings noch etwas mehr. Insbesondere einige der wichtigsten Ergebnisse von TREC wären eine sinnvolle Ergänzung. So haben Relevanz Feedback Strategien sehr großen Erfolg bewiesen. Die Qualität der besten Systeme ist annähernd gleich, jedoch ist die Schnittmenge der Treffer klein.

Kapitel 4 (mit G. Navarro) behandelt Anfragesprachen und diskutiert die Unterschiede zwischen natürlichsprachlichen und booleschen Anfragen, Anfragen mit einem Begriff und Anfragen mit Kontext. Daneben werden strukturierte Anfragen behandelt.

Die Verfahren zur automatischen Verbesserung von Anfragen bespricht ausführlich Kapitel 5. Die Trennung vom Modellkapitel ist durchaus sinnvoll, da Relevanz Feedback, Cluster-Analyse und Assoziationsthesauri unabhängig vom Modell wirken.

Kapitel 6 informiert über Metadaten und anschliessend über Textformate wie Markup-Sprachen und Multimedia-Formate.

Kapitel 7 (N. Ziviani) ist mit Text-Operationen überschrieben und geht zunächst auf lexikalische Analyse, Stoppwörter und Grundformreduktion ein. Dann folgt ein Ausflug zu Thesauri und erneut zu Clustering. Der abschliessende Teil zu Text Kompression hätte durchaus etwas kürzer ausfallen können. Dafür wäre eine detailliertere Beschreibung linguistischer Verfahren ebenso wie die Erwähnung intellektuellen Indexierens angebracht. Dann hätte das Kapitel auch gut vor dem Modell-Kapitel plaziert werden können.

Das Kapitel Indexieren und Suchen (mit G. Navarro) fokussiert auf den effizienten Umgang mit Datenstrukturen im Bereich IR. Neben invertierten Listen werden Suffix-Bäume und Suchalgorithmen abgehandelt.

Kapitel 9 (E. Brown) schliesst nahtlos an Kapitel 8 und führt in die Parallelverarbeitung ein, die eine weitere Effizienzsteigerung verspricht. Unter dem Titel Parallel and Distributed IR vermutet man aber auch neuronale Netzwerkmodelle, die aber hier keine Rolle spielen.

Damit schliesst der technisch orientierte Durchlauf durch IR Themen. Der Aufbau der Kapitel 4, 6, 7 und 8 ist dabei nicht so stringent wie der Rest des Buches. Es folgt das umfangreichste Kapitel des Buches, das Benutzungsoberflächen und Visualisierung im IR behandelt. Erst hier diskutieren die Autoren die Vagheit des Suchprozesses: „Information seeking is an imprecise process“ (S. 257). Das Kapitel beginnt mit einem zu kurzen Abschnitt über Softwareergonomie, der aus der grossen Menge vorhandener Richtlinien drei sehr beliebig herauszieht. Das Thema wird auch nicht ausreichend problematisiert und es entsteht der Eindruck, man müsse nur diese Richtlinien anwenden. Jedoch sind die Richtlinien zum einen sehr schwer umzusetzen und führen zum anderen zu widersprüchlichen Forderungen, so dass im Design-Prozess meist ein ausgewogener Kompromiss gesucht wird.

Der folgende umfassende Überblick diskutiert Benutzungsoberflächen (M. Hearst) und gliedert sich nach Schritten im Suchprozess. Er umfasst die Auswahl des Ausgangspunkts, die Anfrage, Kontext-Darstellungen, Relevanz Feedback und allgemeine Unterstützung des Suchprozesses etwa durch Fenster-Management und Visualisierung des Gesamtablaufs.

Die letzten fünf Kapitel stellen wichtige Anwendungen von IR vor. Modelle und Sprachen für Multimedia IR folgen in Kapitel 11 (E. Bertino, B. Catania, E. Ferrari). Ein solches Kapitel darf in keinem IR Buch mehr fehlen. Die Zugänglichkeit von Multimedia-Daten und der offensichtliche Bedarf nach Retrieval aus heterogenen Quellen verbieten die ausschliessliche Konzentration auf Texte. Gerade da das Kapitel in einer Linie mit der

modernen Forschung liegt, irritiert der fast ausschliessliche Bezug auf das Projekt MULTOS. Dessen Ergebnisse wurden nun schon vor fast zehn Jahren publiziert, so dass sie in diesem dynamischen Feld nicht mehr als alleiniger Maßstab dienen dürfen. Damit ergibt sich auch ein sehr enger Blickwinkel auf Multimedia, der im wesentlichen auf die Integration von DBMS und IR und die Erweiterung von SQL fällt. Diese technische Dimension kann heute schon anhand vorhandener kommerzieller Produkte diskutiert werden.

Kapitel 12 (C. Faloustos) bleibt beim Thema Multimedia und konzentriert sich auf Indexieren und Suchen. Dabei kommen weitere interessante Datentypen wie Zeitreihen und Bilder und ihre inhaltliche Analyse zur Sprache. Interessante neuere Forschungsergebnisse im Bereich Multimedia IR etwa [Schäuble 1997] oder [Dunlop 1997] hätten beiden Kapiteln gut getan.

Kapitel 13 beschäftigt sich mit IR im Internet und führt in die besonderen Herausforderungen ein, wobei insbesondere die enorme Datenmenge, die Mehrsprachigkeit und die Durchdringung mit multimedialen Daten zu Schwierigkeiten für die Standard-Modelle führen. Im folgenden Abschnitt zur Suche haben die bekannten Suchmaschinen des Internet ihren Auftritt. Neben den allzu schnell obsoleten Informationen zu aktuellen Suchmaschinen und Anzahl der (angeblich) indexierten Dokumenten erfährt der Leser auch Grundsätzliches zur Funktionsweise. Suchmaschinen bedienen sich der Hilfe von Crawlern oder Gatherern, die das Netz vollautomatisch absuchen, die riesigen Indizes müssen sehr effizient durchsucht werden und für das Ranking ergeben sich im Internet neben der klassischen inversen Dokument-Frequenz eines Terms und der Frequenz des Terms im Dokument neue Maße wie etwa die Anzahl der Links, die auf eine Seite weisen. Der letzte Abschnitt über Browsing bespricht Internet-Kataloge.

Kapitel 14 (E. Rasmussen) über Bibliotheks- und bibliographische IR Systeme steht vor den digitalen Bibliotheken, es hätte jedoch noch besser an den Anfang der Anwendungen gepasst. Es enthält Hinweise auf kommerzielle Anbieter von Text-Datenbanken und ihre IR Systeme. Ein Überblick über OPACs (Online Public Access Catalogs) leitet über zum letzten, etwas knappen Kapitel über digitale Bibliotheken (E. Fox, O. Sornil). Die nach Definitionen und Architekturen diskutierten Probleme ähneln denen des IR im Internet. Der hohe Grad der Strukturierung in Bibliotheken führt jedoch zusätzlich zum Problem der technischen und semantischen Integration verteilter Datenbestände. Zahlreiche aktuelle, internationale Projekte werden kurz angesprochen.

Trotz genannter Kritik kann der Band insgesamt empfohlen werden. Für die Lehre muss die Dimension der Benutzerorientierung allerdings von Anfang an durch weitere Texte abgedeckt werden. Für Praktiker kann das Buch mit seinem Glossar und Index auch als Nachschlagewerk dienen. Das umfangreiche Literaturverzeichnis und Hinweise auf Forschungsthemen mit Referenzen am Ende jedes Kapitels bieten für alle Themen die Möglichkeit zur Vertiefung.

(Thomas Mandl, Universität Hildesheim, Informationwissenschaft,
mandl@rz.uni-hildesheim.de)

Schäuble, Peter (1997): Multimedia Information Retrieval: Content-Based Information Retrieval from Large Text and Audio Databases.

Dunlop, Mark (1997): Proceedings of the Second Mira Workshop. Monselice, Italien, November 1996.