Vorwort

Die Suche nach Texten zu einem bestimmten Thema hat sich durch das World Wide Web in den letzten Jahren von einer Aufgabe in (wissenschaftlichen) Bibliotheken und Sammlungen zu einem alltäglichen Problem vieler Menschen entwickelt. Im Web gefunden zu werden, kann – nicht nur für Unternehmen – ein entscheidender Erfolgsfaktor sein.

Information Retrieval (IR) als wissenschaftliche Disziplin, die die inhaltliche Suche nach Informationen in Sammlungen von »Dokumenten« untersucht und Modelle, Methoden und Verfahren dafür entwickelt, hat dadurch aber nicht entsprechend größere Beachtung gefunden. Häufig werden bei der Entwicklung des Web eher einzelne Technologien und Dienste wahrgenommen als eine zusammenfassende Sicht aus der Perspektive der inhaltlichen Suche.

Das vorliegenden Buch versucht eine integrierte Darstellung des IR zu geben, die von den klassischen Methoden wie Klassifikationen und Thesauren bis zur Suche im WWW reicht. Schwerpunkte liegen dabei auf der Darstellung der Bezüge zu anderen Disziplinen und auf Verfahren, mit denen Wissen aus Sammlungen gewonnen werden kann, um die Suche zu unterstützen.

Die Darstellung konzentriert sich auf Modelle und Methoden der Suche nach Textdokumenten, auch wenn einige der Modelle auf andere Informationsarten übertragen werden können. Sie versucht die Anwendung im Auge zu behalten, ohne dabei zu sehr in technische Details zu gehen oder Rezepte anzubieten. Auf die Darstellung konkreter Implementierungen und Systeme wurde in der Regel zugunsten der konzeptionellen Sicht verzichtet. Teilweise werden Experimente und deren Ergebnisse genauer beschrieben, um aktuelle Entwicklungen und deren Komplexität darzustellen.

Das Buch richtet sich an Studierende und Berufstätige, die sich die Grundlagen des IR aneignen wollen. Darüber hinaus führt es in moderne Methoden und Verfahren – insbesondere auch aus der Web-Suche – ein, beschreibt modellhafte Beispiele, stellt sie in einen theoretischen Zusammenhang und unterstützt damit den Zugang zur aktuellen IR-Forschung.

Ziel des Buchs ist es, seinen Leserinnen und Lesern ein solides Grundlagenwissen in Information Retrieval zu vermitteln und dessen Stellung zwischen Ingenieur- und Humanwissenschaft deutlich zu machen. Darüber hinaus soll es sie in die Lage versetzen, Inhalts- und Suchmodelle und -systeme, Entwicklungen und Trends zu verstehen, die verwendeten Methoden zu erkennen und abzuschätzen, ob sie sinnvoll eingesetzt werden. Es hat nicht den Anspruch, eine Anleitung zum Bau einer Suchmaschine zu sein.

Die Lektüre setzt an einigen Stellen etwas mathematisches Verständnis voraus, wobei alle wichtigen Begriffe und Konzepte eingeführt werden. Das gilt auch für Konzepte aus Nachbardisziplinen wie unscharfe Mengen, Wahrscheinlichkeitsrechnung oder Begriffe aus der Lerntheorie. Einige der mathematisch detaillierter dargestellten Passagen können übersprungen werden, ohne dass dadurch das Verständnis für andere Teile des Buchs behindert wird.

Das Buch ist in vier Teile gegliedert:

Der erste Teil führt mit Beispielen und einigen theoretischen Überlegungen ins Thema ein und beschreibt die klassischen Methoden und Hilfsmittel der Dokumenterschließung und -suche, wie Klassifikationen, Thesauren, boolesche Suche und das Vektorraummodell. Weiter werden Verfahren vorgestellt, mit denen Suchsysteme bewertet werden können.

Der zweite Teil gibt eine Einführung in die Wissensgewinnung mit Data-Mining-Methoden, also das »Lernen« aus Beispielen und Sammlungen. Er stellt verschiedene Ansätze und Verfahren vor, wie Entscheidungsbäume und Regelsysteme, diskutiert die Rahmenbedingungen ihres Einsatzes und beschreibt eine konkrete Anwendung.

Diese ersten beiden Teile können als Grundkurse für das jeweilige Gebiet genutzt werden. Sie enthalten einige vertiefende Abschnitte, die gegebenenfalls übersprungen werden können.

Im dritten Teil werden moderne Entwicklungen im Information Retrieval und Verfahren beschrieben, die Wissensgewinnungsmethoden für das IR nutzen. Dieser Teil setzt die beiden ersten Teile voraus. Er zeigt, welche neuen Ansätze in den letzten Jahren entwickelt wurden, gibt Einblick in die Komplexität der verwendeten Verfahren und dient als Brücke zu Studium und Verständnis der aktuellen IR-Forschung.

Der vierte und letzte Teil des Buchs widmet sich der Anwendung von IR-Verfahren im World Wide Web. Die dort verwendeten Auszeichnungs- und Repräsentationsmethoden wie HTML, XML, RDF und Metadaten-Systeme werden ebenso beschrieben wie die Rahmenbedingungen, Methoden und Perspektiven für die Suche im Web. Dieser Teil setzt im Wesentlichen nur den ersten Teil des Buchs voraus.

Das vorliegende Buch ist aus den Skripten zu zwei Vorlesungen entstanden, die ich zwischen 1995 und 2000 mehrfach am Fachbereich Informatik der Technischen Universität Darmstadt zu den Themen »Data Mining und Information Retrieval« und »Informationssysteme« gehalten habe.

Bei dem Vorhaben, aus den Vorlesungsskripten ein Buch zu machen, bin ich von verschiedenen Seiten unterstützt worden: von zahlreichen Hörerinnen und Hörern der Vorlesungen durch Diskussionen, Rückmeldung und konstruktive Kritik, von Ute Sotnik durch die Korrektur der ersten Version des Manuskripts, von Eva Emskötter durch ausdauernde Hilfe beim Erstellen der endgültigen Version, von den Mitarbeitern und Mitarbeiterinnen des dpunkt-Verlags durch gute Betreuung und Zusammenarbeit.

Dafür bedanke ich mich herzlich.

Münster in Westfalen im Februar 2003
Reginald Ferber