1 Einführende Beispiele

Die zunehmende Verbreitung von elektronischen Informationsverarbeitungs- und Speichermedien lässt die Menge der Daten, Texte, Bilder und Videofilme, die digital zur Verfügung stehen, immer schneller anwachsen. Gleichzeitig werden diese Daten durch die zunehmende Vernetzung für immer mehr Menschen zugänglich. Das gilt sowohl innerhalb einzelner Organisationen als auch weltweit zwischen Organisationen und Einzelpersonen.

Diese Daten können aber nur genutzt werden, wenn sie auch erschlossen sind, wenn also diejenigen, die sie nutzen wollen oder sollen, auch wissen, dass und wo es die Daten gibt, wie sie gesuchte Informationen darin finden können und wie sie diese nutzen können und dürfen. Um zu illustrieren, wie unterschiedlich die Bedingungen sein können, kann man die klassische Telefonauskunft und das Web als Beispiele für Systeme betrachten, die Informationen zur Verfügung stellen.

Die Telefonauskunft setzt voraus, dass die Nutzenden wissen, wozu ein Telefon gut ist, wie man damit umgeht und wessen Telefonnummer sie suchen. Die gegebene Information hat einen genau bestimmten Zweck: den entsprechenden Telefonanschluss zu erreichen. Wenn die Telefonnummer zu einem bestimmten Namen erfragt wird, werden die Angaben von den Nutzenden im Allgemeinen als verlässlich angesehen, da kein unmittelbarer Grund ersichtlich ist, warum eine falsche Auskunft erteilt werden sollte. (Wenn nicht nach einer Adresse, sondern nach einer Dienstleistung gefragt wird – wie der Nummer eines Rechtsanwalts oder einer Ärztin – sieht das schon anders aus.)

Beim *World Wide Web* (im Folgenden auch einfach *Web* genannt) sind die Verhältnisse weniger klar. Es besteht aus Milliarden Dateien, die weltweit auf unzähligen Rechnern verteilt sind und deren wesentliche Gemeinsamkeiten das Übertragungsprotokoll HTTP (Hypertext Transfer Protocol) und zum Teil die Auszeichnungssprache HTML (Hypertext Markup Language) sind. Weder die Inhalte noch der Zweck des Systems sind klar definiert. Entsprechend aufwändiger ist es hier, nach bestimmten Inhalten zu suchen. Hat man Informationen gefunden, ist es schwierig zu beurteilen, ob sie richtig, vollständig und zuverlässig sind.

Während die klassische Telefonauskunft also einen sehr spezifischen Informationsbedarf in einem genau festgelegten Handlungsablauf bedient, waren beim Web zunächst weder für die Inhalte, noch für die Art, in der mit ihnen umgegangen wird, feste Regeln vorgegeben. In den ersten Jahren schien es eher durch die technischen Möglichkeiten als durch einen tatsächlichen In-

formationsbedarf bestimmt zu sein. Erst im Lauf der Zeit hat sich das Web zu einem wichtigen Informationsmedium entwickelt, das auch durch entsprechende Gesetze und wirtschaftliche Erwartungen reglementiert ist. Dabei bilden sich aber vor allem einzelne kontrollierte Angebote innerhalb eines unkoordinierten Gesamtsystems heraus.

Das Problem, in wenig koordinierten und kontrollierten »Sammlungen« die »richtigen« Informationen oder Dokumente zu finden, ist allerdings weder neu, noch auf elektronische Medien beschränkt. So wurde z. B. im Bereich des (wissenschaftlichen) Literatur- und Bibliothekswesens schon lange versucht, Systematiken zu entwickeln, nach denen Artikel und Bücher inhaltlich erfasst, bewertet und geordnet werden können. Die dafür entwickelten Systeme waren aber meist auf vergleichsweise kleine oder wohl definierte Dokumentensammlungen, Sachgebiete und Gruppen von Nutzenden beschränkt. Sie stützen sich häufig auf die Begutachtung und Einordnung der Dokumente durch Fachleute.

Dadurch, dass die Digitalisierung und Vernetzung zunimmt und dass es für viele Menschen immer einfacher und billiger wird, selbst Dokumente zu erstellen und öffentlich zugänglich zu machen, stellen sich viele Probleme neu.

Die Faktoren Informationsbedarf und technische Machbarkeit haben bei der Entwicklung von Informationssystemen natürlich immer eine Rolle gespielt: Zum einen muss ein hinreichend großer Informationsbedarf vorhanden sein, damit ein System entwickelt wird, zum anderen setzen die technischen Möglichkeiten häufig Grenzen. Historisch sind beide Faktoren für elektronische Systeme wesentlich wichtiger gewesen, als sie heute erscheinen. Computer waren noch vor 30 Jahren sehr teuer und vor allem selten. Sie standen fast nur großen wissenschaftlichen Einrichtungen zur Verfügung. Ihr Einsatz musste explizit gerechtfertigt und genehmigt werden. Zudem waren sie im Vergleich zu heute nicht sehr weit entwickelt. Insbesondere das Speichern großer Datenmengen war aufwändig und teuer.

In dieser Situation wurden erste Systeme entwickelt, um den ständig wachsenden Bestand an (wissenschaftlicher) Literatur zu verwalten und nach inhaltlichen Kriterien darauf zugreifen zu können. Dazu entstanden zunächst Katalogsysteme, die zunehmend auch inhaltliche Beschreibungen der katalogisierten Dokumente enthielten. Aus diesem Ansatz hat sich das Fachgebiet *Information Retrieval* (*IR*) entwickelt. Es beschäftigte sich zunächst vor allem damit, die relevante Literatur zu einer wissenschaftlichen Fragestellung zu finden. (Aus diesem Grund bezeichnet man im Allgemeinen die Objekte, nach denen mit IR-Systemen gesucht wird, als *Dokumente*, selbst wenn es sich nicht um Texte handelt.)

Im Folgenden werden verschiedene Möglichkeiten, Literatur zu einem Thema zu suchen, kurz dargestellt. Anschließend werden exemplarisch einige Beispiele für Systeme, die Informationen vermitteln, beschrieben.

1.1 Literatursuche

Um Literatur zu einem wissenschaftlichen Thema zu finden, gibt es eine ganze Reihe verschiedener Herangehensweisen. Man kann jemanden *fragen*, der oder die sich auskennt, ein *Buch* suchen, das eine Einführung in das fragliche Gebiet gibt, *Literaturverweise* in Büchern und Artikeln weiterverfolgen, in einer thematisch geordneten *Bibliografie* oder *Abstract-Sammlung* nachsehen, in einer *elektronischen Literaturdatenbank* suchen, oder mit Hilfe von *Übersichtsseiten* oder *Suchmaschinen im Web* nach Material suchen. Die verschiedenen Vorgehensweisen haben jeweils Vor- und Nachteile:

Befragen von Expertinnen und Experten

Um eine *Expertin* oder einen *Experten* um Rat zu fragen, muss man erst eine Person finden, die sich auskennt und Zeit und Lust hat, sich mit der Frage zu beschäftigen. Das ist häufig dann schwierig, wenn man neu in einem Gebiet ist, oder Fragestellungen interdisziplinär sind. Hat man eine solche Person gefunden, besteht allerdings die Möglichkeit, die Fragestellung im Gespräch weiter zu erörtern und dabei zu präzisieren. Es besteht aber auch die Gefahr, dass das Problem nur der Sichtweise und dem Kenntnisstand der befragten Person entsprechend angegangen wird.

Bücher und Tagungsbände

Bücher erscheinen häufig erst, wenn ein neues Fachgebiet oder eine neue Sichtweise schon eine gewisse Zeit existiert. Sie sind langsam. *Konferenzbände* enthalten Artikel zu Vorträgen, die auf einer wissenschaftlichen Konferenz gehalten wurden. Sie sind daher häufig aktuellere Zusammenstellungen von Artikeln zu einem Thema als Bücher. Sie bieten allerdings keine geschlossene und systematische Darstellung eines Gebiets, da sich die einzelnen Beiträge nicht aufeinander beziehen und für jeden Beitrag nur beschränkter Raum zur Verfügung steht. Außerdem ist die Auswahl der Artikel nicht immer nur themen- oder qualitätsbezogen. Häufig spielen bei der Zusammenstellung von Beiträgen zu einer wissenschaftlichen Konferenz andere Gesichtspunkte eine Rolle.

Bibliotheksrecherchen

Mit den klassischen Methoden der *Papierbibliothek* dauerte das Beschaffen von Literatur häufig lange. Wenn nur der Titel eines Artikels bekannt ist, lässt sich zusätzlich oft schwer einschätzen, ob er eine Fragestellung abdeckt. Dadurch kann es auch nötig werden, mehrmals nacheinander Artikel zu beschaffen, was eine weitere Verzögerung bedeuten kann. In den letzten Jahren sind allerdings elektronische Bestell- und Lieferdienste entwickelt worden, die auch die Lieferung von Kopien von »Papierartikeln« (per Fax oder als eingescanntes Bild) erheblich beschleunigt haben.

Stichwortkataloge und *Inhaltsklassifikationen* sind hierarchisch nach *Themengebieten* aufgebaut. Ihre Verwendung setzt daher Kenntnisse über die Sachgruppenhierarchie (und damit das Gebiet der Fragestellung) voraus. Schwierige Suchprobleme zeichnen sich häufig gerade dadurch aus, dass die Fragestellungen nicht innerhalb einer etablierten Theorie bleiben oder dass die Suchenden sich in dem Fachgebiet, in dem sie suchen, (noch) nicht gut auskennen. Dadurch können die Anfragen nur schwer auf die vorgegebene Sachgruppenhierarchie abgebildet werden.

Literaturdatenbanken sind häufig noch komplex zu bedienen, teilweise teuer und stehen nicht in allen Bibliotheken zur Verfügung. Sie vermitteln mit Hilfe von Sachgruppen, Stichwörtern und Freitextsuche Informationen über Artikel und Bücher, liefern jedoch häufig nur die bibliografischen Angaben, sodass die Vollversion eines Artikels nach wie vor beschafft werden muss. Das wird heute allerdings zunehmend durch Online-Lieferdienste direkt möglich.

Im Web suchen

Übersichtsseiten im Web sind häufig von einzelnen Personen ehrenamtlich zusammengestellt. Sie weisen damit ähnliche Probleme auf wie die persönliche Nachfrage. Darüber hinaus sind sie meist statisch: Das heißt, es ist selten möglich, Fragen zu stellen (und Antworten auf Fragen zu bekommen) oder einen persönlichen Eindruck von der Person zu gewinnen, auf deren Expertise man sich verlässt. Zunehmend trifft man auch auf Seiten, die seit langem nicht mehr aktualisiert worden sind. Große Themenseiten müssen – wie Bibliotheken oder Abstract-Sammlungen – nach einem inhaltlichen System strukturiert werden. Der Zugang kann für Unerfahrene daher ähnlich schwierig sein wie bei Sachgruppen-Klassifikationen in Bibliotheken.

Web-Suchmaschinen beschränken sich häufig auf die Suche nach einzelnen Wörtern oder Wortgruppen im Text der Dokumente. Bei der enormen Anzahl von Dokumenten zu den unterschiedlichsten Themen ist es mit diesen Mitteln schwierig, vollständige Suchergebnisse zu erzielen. Lange Zeit haben viele Suchmaschinen auch nur HTML-Dokumente verarbeitet, sodass Dokumente in anderen Formaten, wie PostScript oder PDF, nicht erfasst wurden. Schließlich werden zunehmend Web-Angebote in Datenbanken abgelegt, auf deren Dokumente Suchmaschinen häufig nicht zugreifen. Da das Web öffentlich ist, sind Artikel, die verkauft werden sollen (wie z. B. viele Artikel, die in wissenschaftlichen Zeitschriften erscheinen) oder aus anderen Gründen nicht frei zugänglich sind, dort oft nicht zu finden. Schließlich kann die Richtigkeit von Angaben im Web nur bedingt überprüft werden, und es ist auch oft nicht sonderlich schwierig, z. B. eine falsche Urheberschaft einer Information vorzutäuschen.

Nach dieser kurzen Einführung in die Probleme, die bei der Literatursuche auftreten können, werden im Folgenden verschiedene Systeme, mit denen Informationen angeboten werden, vorgestellt, um einen ersten Überblick über unterschiedliche Ansätze zu geben.

1.2 Recherche in einer Literaturdatenbank

Will man beispielsweise Literatur zum Stand der Forschung im Bereich Retrieval-Systeme für Multimedia-Objekte mit einem besonderen Schwerpunkt auf der Frage suchen, wie Bilder behandelt werden, könnte man die Datenbank INSPEC verwenden. Sie besteht aus Dokumenten, die Artikel und Bücher beschreiben, indem sie neben bibliografischen Angaben eine *Kurzzusammenfassung* (*Abstract* oder auch *Referat*), eine Einordnung in ein hierarchisches Klassifikationssystem und Stichwörter enthalten.

Für andere Fachgebiete und Sprachen gibt es andere Literaturdatenbanken. Abbildung 1.1 zeigt einen Eintrag aus der psychologischen Literaturdatenbank PSYNDEX, die vom Zentrum für psychologische Information und Dokumentation (ZPID) an der Uni Trier erstellt und angeboten wird. In diesem Eintrag wird ein Zeitschriftenartikel beschrieben. Neben detaillierten bibliografischen Angaben, wie Autoren, Titel, Erscheinungsjahr und Anzahl der Literaturverweise (Referenzen), enthält der Eintrag eine ausführliche Beschreibung des Inhalts und fast zehn verschiedene Felder zur systematischen inhaltlichen Erfassung. Insgesamt gibt es in PSYNDEX 52 Feldbezeichner, aus denen Beschreibungsformate zusammengestellt werden können. In Abbildung 1.1 sind die ausgeschriebenen Namen der verwendeten Felder in Klammern hinter den Kürzeln aufgeführt.

Die Suche in solchen Literaturdatenbanken geht folgendermaßen vonstatten: Auf die Eingabe von (Kombinationen von) Wörtern liefert das (boolesche) Retrieval-System von INSPEC die Dokumente, die die Wörter in der angegebenen Form enthalten (Genaueres siehe Abschnitt 3.1). Um ein gutes Suchergebnis zu erzielen, sollte man deshalb zunächst überlegen, welche Stichwörter das Problem besonders gut beschreiben. Es sollten Wörter sein, die spezifisch für die Fragestellung sind, aber doch wieder so allgemein, dass man annehmen kann, dass sie in jedem »wichtigen« Artikel vorkommen.

Eine Anfrage mit den Stichwörtern *Retrieval Systems*, *Multimedia* und *Images* könnte z. B. so aussehen:

```
retrieval systems AND multimedia AND images.
```

Sie wird vom Retrieval-System so interpretiert: Suche alle Dokumente, in denen jede der drei Zeichenketten `retrieval systems`, `multimedia` und `images` mindestens einmal irgendwo im Text vorkommt. Dabei wird nicht zwischen Groß- und Kleinbuchstaben unterschieden, und das Leerzeichen zwischen `retrieval` und `systems` kann auch ein anderer »*white space*« sein (also z. B. ein Zeilenumbruch, auch in Verbindung mit mehreren Leerzeichen).

Die Suche in der Literaturdatenbank INSPEC lieferte für die Zeit von Januar bis Juni 1995 drei Einträge: »*Image Engine: an object-oriented multimedia database for storing, retrieving and sharing medical images and text*«, »*Multimedia information retrieval using knowledge in encyclopedia texts*«, »*Images database management system: a 'server-client producer' system on a local network and on the Internet*«.

AN *(Datensatznummer)*: 0153200
DT *(Dokumenttyp)*: Journal-Article (10)
TI *(Titel)*: Bibliometrische Untersuchungsbefunde zur Geschichte der Klinischen Psychologie im 20. Jahrhundert Bibliometric findings on the history of clinical psychology in the 20th century
AU *(Autoren)*: Krampen,-Guenter; Miller,-Marianne; Montada,-Leo
AF *(Institutionelle Zugehörigkeit des Erstautors)*: Universitaet Trier; Fachbereich I - Psychologie, Germany
CY *(Land)*: Germany (D)
EM *(E-Mail-Adressen von Autoren)*: Krampen, Guenter: krampen@uni-trier.de
SO *(Quelle bei Zeitschriftenaufsätzen)*:
Zeitschrift-fuer-Klinische-Psychologie-und-Psychotherapie. 2002; 31(2): 121-126
URL: http://www.hogrefe.de/Zeitschriften/index.html
PY *(Veröffentlichungsjahr)*: 2002
RN *(Anzahl der Referenzen)*: 20
IS *(ISSN, International Standard Serial Number)*: 1616-3443
LA *(Sprache)*: German
AL *(Sprache des Kurzreferats)*: German
ABG *(Deutsches Kurzreferat)*: Unter fachhistoriographischer Orientierung werden bibliometrische Untersuchungsbefunde zur Entwicklung des Fachliteraturaufkommens zur Klinischen Psychologie im 20. Jahrhundert praesentiert. Im Mittelpunkt steht dabei die Frage, wie die Geschichte der Klinischen Psychologie und ihrer Teilbereiche im 20. Jahrhundert bibliometrisch, (...), rekonstruiert werden kann. Als Datenbasis dienten die Psychological Abstracts (1927-1966) sowie die Literaturdatenbanken PsycLit (1967-1999) und PSYNDEX (1977-1999), in denen die psychologische Fachliteratur aus dem internationalen (primaer angloamerikanischen) bzw. deutschsprachigen Bereich relativ exhaustiv dokumentiert wird. (...) Auf Probleme bibliometrischer Analysen wird ebenso verwiesen wie auf Implikationen der Befunde fuer die Zukunft der Klinischen Psychologie, insbesondere der Psychotherapieforschung. (Zeitschrift/U.R.W.ZPID)
KP *(Knappe Inhaltsbeschreibung in englischer Sprache)*: bibliometric findings on history of clinical psychology in 20th century; boom since 1950s & percentage of publications on psychotherapy & frequency of empirical & experimental methods; analysis of Psychological Abstracts & literature databases PsycLit & PSYNDEX; bibliometric study
MJ *(Englische Hauptschlagwörter)*: *Clinical-Psychology; *History-of-Psychology
MN *(Englische Nebenschlagwörter)*: Scientific-Communication; Psychotherapy-; Empirical-Methods; Experimental-Methods; Trends-
MJG *(Deutsche Hauptschlagwörter)*: *Klinische-Psychologie; *Geschichte-der-Psychologie
MNG *(Deutsche Nebenschlagwörter)*: Wissenschaftliche-Kommunikation; Psychotherapie-; Empirische-Methoden-; Experimentelle-Methoden; Trends-
CX *(Englische Sachgebiete)*: Psychological-and-Physical-Disorders; Health-and-Mental-Health-Treatment-and-Prevention
CG *(Deutsche Sachgebiete)*: Psychische-und-physische-Stoerungen; Behandlung-und-Praevention
CC *(Sachgebiets-Codes)*: 3200; 3300; 32; 33
PT *(Publikationstyp)*: empirical-study
PTC *(Publikationstyp-Code)*: 1010; 10
UD *(Eingabedatum)*: 200207

Abbildung 1.1 – Dokument aus der Literaturdatenbank PSYNDEX: Vor dem (hier aus Platzmangel um mehr als die Hälfte gekürzten) Abstract oder Kurzreferat stehen im Wesentlichen bibliografische Angaben. Danach folgen Felder, in denen der Inhalt des Artikels nach verschiedenen Systematiken beschrieben wird. (Abdruck mit freundlicher Genehmigung des ZPID, Universität Trier, 2002)

Anfrage	Treffer
`retrieval systems AND multimedia AND images`	3
`retrieval systems AND multimedia AND image$`	5
`retrieval AND multimedia AND image$`	35
`retrieval AND multimedia`	148
`retrieval OR multimedia`	2559
`retrieval OR multimedia OR image$`	9364

Abbildung 1.2 – Anzahl der in INSPEC gefundenen Dokumente im ersten Halbjahr 1995: In der Anfrage sind AND und OR boolesche Operatoren, mit denen die Anfrageterme verknüpft werden. Sind zwei Suchterme mit AND verknüpft, wird ein Dokument ausgegeben, wenn es beide Suchterme enthält; sind sie mit OR verknüpft, wird ein Dokument ausgegegeben, wenn es den einen oder den anderen Term enthält (oder beide). Die Ergebnismenge bei einer OR-Verknüpfung enthält also die Ergebnismenge als Teilmenge, die sich bei einer AND-Verknüpfung ergibt.

Nun sind drei Dokumente nicht gerade viel. Man kann also versuchen, die Anfrage etwas allgemeiner zu formulieren. Wenn die folgende Anfrage verwendet wird: `retrieval AND multimedia AND image$`, finden sich immerhin 35 Dokumente für denselben Zeitraum. Dabei ist `$` eine *Wildcard*, `image$` bezeichnet also alle Wörter, die mit der Zeichenkette `image` beginnen. In der Ergebnisliste finden sich Titel wie: »*PhotoFile: a digital library for image retrieval*«, » *Spatial knowledge representation and retrieval in 3-D image databases*«, »*Multimedia retrieval technology*«, »*A WWW interface to the OMNIS/Myriad literature retrieval engine*«, »*Problems of content-based retrieval in image databases*«. Sie sehen auf den ersten Blick nicht weniger relevant aus als die oben genannten. Weitere Verallgemeinerungen der Anfrage und ihre Treffermengen sind in Abbildung 1.2 angegeben.

Zu den Titeln aus der Ergebnisliste können noch die vollständigen bibliografischen Angaben abgerufen werden, ähnlich wie sie in Abbildung 1.1 für die psychologische Datenbank gezeigt wurden. Sie enthalten auch ein Abstract und Einordnungen in verschiedene Ordnungs- und Beschreibungssysteme. Die vollständigen Artikel müssen aber auf anderen Wegen, also z. B. aus der Bibliothek, über die Fernleihe oder durch einen Lieferdienst, beschafft werden. Neuere Systeme bieten auch die Möglichkeit, die vollständigen Artikel elektronisch zu bestellen oder unmittelbar als Volltextdatei zu beziehen. Die Probleme, die sich dabei ergeben, sind im Allgemeinen eher ökonomischer oder organisatorischer Natur als technischer.

1.3 Faktendatenbanken und -retrieval

Während die Dokumente in Literaturdatenbanken zwar in verschiedene Felder strukturiert sind, innerhalb vieler dieser Felder aber freien Text beliebiger Länge enthalten, sind die Einträge in Faktendatenbanken im Allgemeinen stark strukturiert. Das heißt, sie bestehen (logisch) aus Tupeln von Werten, für die bekannt ist, welchen Datentyp sie haben – ob es sich also beispielsweise um Zahlen, Zeitangaben, Preise, Eigennamen oder freie Text handelt. Abbildung 1.3 zeigt eine fiktive Beispieldatenbank mit unterschiedlichen Datentypen. Faktendatenbanken werden im Allgemeinen mit relationalen Datenbankmanagementsystemen (DBMS) verwaltet. Diese Systeme sorgen neben der Suche vor allem für die Konsistenz und Sicherheit der verwalteten Daten, auch wenn sie von mehreren Nutzenden gleichzeitig bearbeitet – und insbesondere geändert – werden. Dieser Aspekt wird von IR-Systemen in der Regel nicht berücksichtigt und soll hier auch nicht näher betrachtet werden (siehe z. B. Grossman und Frieder, 1998 [49], Kapitel 5).

	m^2	Kaltmiete	Zimmer	Balkon	Ort	Stockwerk	Heizung
A	64	420	3 ZKB	n	Kranichstein	13	zentral
B	78	650	4 ZKB	j	Bessungen	2	Gasetage
C	86	775	3 ZKB	j	Martinsviertel	4	zentral Fußboden
D	102	310	3	n	Wiebelsbach	EG	Ofen
E	36	380	2 ZKB	j	DA-Ost	3	Nachtspeicher
F	34	340	3 ZKB	n	Arheilgen	EG	Öl
G	38	290	1,5 ZB	j	Griesheim	2	zentral
H	87	490	4 ZKB	n	Heimstättensiedlung	3	zentral

Abbildung 1.3 – Beispieldatenbank mit Wohnungsangeboten: Die Spalten enthalten Werte unterschiedlichen Typs: In der 2. und 3. Spalte stehen reelle Zahlen, die nach ihrer Größe verglichen werden können. Eine Bewertung wird im Allgemeinen ergeben, dass in der 2. Spalte größere Werte besser sind, in der 3. aber kleinere. Bei den übrigen Spalten ist sowohl der Vergleich als auch die Bewertung schwieriger. So kann es sein, dass für das Stockwerk ein Wert zwischen 2 und 5 als Optimum angesehen wird, also keine monotone Bewertung stattfindet.
Weiter zeigt sich, dass einige der Angaben in Verbindung mit anderen Werten bzw. mit Weltwissen aussagekräftiger werden.

Die starke Strukturierung der Datensätze erleichtert den Zugriff auf die Einträge und das Arbeiten mit ihnen, da die Typisierung das Format vorgibt und wohl definierte Vergleiche ermöglicht. Diese Stärke kommt vor allem zum Tragen, wenn exakte Anfragen gestellt werden, wie – um bei der Beispieldatenbank zu bleiben – *»suche eine Wohnung mit einer Quadratmeterzahl zwi-*

schen 65 und 85 und einer Kaltmiete unter 500 Euro«. Wenn die Anfragen allerdings vager werden, müssen die Werte häufig wieder interpretiert werden, wenn für Anfragende nützliche Ergebnisse erzielt werden sollen. So kann man versuchen, einen Informationsbedarf wie *»Suche stadtnahe, kostengünstige Wohnung für zwei Personen«* in eine Anfrage mit den gegebenen Attributen zu übersetzen. Dazu ist es aber notwendig, weiteres Wissen über das Gebiet zu haben, aus dem die Datensätze sind, und dieses auch in geeigneter Weise nutzen zu können.

Nicht nur durch einen vagen Informationsbedarf können Probleme entstehen. Es kann auch vorkommen, dass die gewünschte Information nicht in der Darstellung der Objekte in der Datenbank vorhanden ist oder eventuell überhaupt nicht geeignet beschrieben werden kann. So lässt sich in einer Faktendatenbank über Bücher gegebenenfalls einfach feststellen, dass ein Buch mit 200 Seiten mehr Seiten hat als eines mit 150 (weil der Datentyp Ordinalskalenniveau hat – oder anders gesagt, weil man für zwei verschiedene natürliche Zahlen immer sagen kann, welche größer ist). Dagegen kann es erheblich schwieriger zu entscheiden sein, welches von zwei Büchern sich besser dazu eignet, sich in ein bestimmtes Thema einzuarbeiten (weil dafür keine Attribute angegeben sind oder weil für die entsprechenden Attributwerte kein offensichtlicher Vergleichsoperator existiert). Andererseits wird in vielen Fällen die zweite Art von Information nützlicher sein als die erste, wenn es darum geht, sich zwischen zwei Büchern zu entscheiden.

1.4 Hypertext-Informationssysteme

Seit der Verbreitung des World Wide Web werden dort Informationsangebote für Organisationen wie Universitäten, Städte oder Unternehmen und Ereignisse wie Kongresse oder Messen angeboten: Von einer *Startseite* (Homepage), die als Web-Adresse angegeben wird, sollen alle Informationen über die Organisation oder das Ereignis erreichbar sein. Die klassische Struktur solcher Angebote ist ein Hypertext-System, also ein System von Seiten, die auf vorgegebenen »Pfaden« aus Verweisen (Links) von der Startseite aus erreicht werden können. Auch wenn bei größeren Angeboten heute häufig weitere Zugriffsmethoden wie eine lokale Suchfunktion hinzukommen, ist die Verlinkung der Seiten eines Web-Angebots immer noch der wichtigste Zugriffsweg. Für ein Hypertext-System müssen die Informationen so aufbereitet und gegliedert werden, dass Nutzende bereits auf der Startseite – also bei einer sehr groben Einteilung – entscheiden können, auf welchem Pfad oder unter welcher Rubrik sie die Information, die sie suchen, finden können.

Auf der Homepage der Stadt Darmstadt fanden sich 1998 z. B. sechs Rubriken mit den Bezeichnungen

- Städtische Einrichtungen
- Kunst & Kultur
- Zu Gast in Darmstadt
- Darmstädter Leben
- Wirtschaft
- Darmstadt aktuell

Sucht man in einer solchen Auswahl z. B. Informationen über Hotels, wird man sicherlich zunächst die Rubrik *Zu Gast in Darmstadt* wählen. Dort fand sich dann eine Rubrik *Hotels und Restaurants*, die weiterführte zu *Hotels in der Innenstadt*, *Hotels in den Vororten* etc. Unter *Hotels in der Innenstadt* gelangte man zu einer Liste mit Adressen, aus der teilweise auch auf Seiten der einzelnen Häuser verwiesen wurde.

Wollte man aber etwas über die Parks in Darmstadt erfahren, war zunächst nicht klar, unter welcher Rubrik man nachsehen sollte. Unter *Städtische Einrichtungen* fand sich das Gartenamt, dort wurden aber nur die Öffnungszeiten des Amtes angegeben. Unter *Darmstädter Leben* fand sich die Seite *Sport und Freizeit* mit einer allgemeinen Erwähnung der Parks. Unter *Zu Gast in Darmstadt* schließlich konnte man unter der Rubrik *Virtueller Stadtrundgang* Seiten zum Herrngarten, zum Prinz-Georgs-Garten, zur Rosenhöhe und zum Platanenhain finden, also zu den wichtigsten Parks der Stadt (wenn man vom Bürgerpark-Nord mit seinen legendären Mitternachtsflohmärkten absieht).

Diese Beobachtung beleuchtet ein generelles Problem von Hypertext-Systemen, an dem sich auch in den letzten Jahren nichts geändert hat: Sie müssen die Balance halten zwischen einer klaren Strukturierung der angebotenen Information, die dann aber immer nur eine Sichtweise widerspiegeln kann, und einer möglichst umfassenden Vernetzung, die aber schnell dazu führt, dass sich Nutzende nicht mehr zurecht finden und den so genannten »Lost-in-Hyperspace-Effekt« erleben: das orientierungslose Herumspringen zwischen den Seiten eines Hypertext-Systems. Dieser Effekt wird als besonders ärgerlich empfunden, wenn man eine Seite schon einmal gesehen hat, sie aber nicht wieder finden kann.

1.5 Expertensysteme

Bei den bisher besprochenen Beispielen unterschieden sich im Wesentlichen die Zugriffsmethoden auf gespeicherte Informationen voneinander. Bei Expertensystemen sind die Informationen, die den Nutzenden angeboten werden, selbst nicht mehr fest gespeichert, sondern werden für jede Anfrage aus zugrunde liegenden Wissensbasen neu generiert.

Als Beispiel kann ein Fahrplan-Auskunftssystem dienen. Hier lassen sich nicht mehr alle möglichen Verbindungen zwischen beliebigen Bahnhöfen einzeln in der Form speichern, in der sie auch ausgegeben werden. Statt dessen wird die Antwort des Systems berechnet bzw. aus dem vorhandenen Wissen hergeleitet. Wenn nach einer Verbindung zwischen zwei Bahnhöfen gefragt wird, kann z. B. zunächst festgestellt werden, welche möglichen Strecken es zwischen den Bahnhöfen gibt. Aus den möglichen Strecken müssen diejenigen herausgesucht werden, die entweder die kürzeste, die schnellste oder die bequemste Verbindung erlauben (oder die, bei der die Bahn am meisten verdient). Andere Randbedingungen können bestimmte Zugtypen sein oder Zeiten, in denen eine Strecke nicht befahren wird. Um eine solche Berechnung durchzuführen, müssen die Daten im Allgemeinen in genau spezifizierten Formaten und Typen vorliegen, auf denen die Systeme arbeiten können.

Bei Fahrplan-Auskunftssystemen kann es auch sinnvoll sein, für eine eingegebene Adresse zunächst zu bestimmen, welche öffentlichen Verkehrsmittel am einfachsten zu erreichen sind. Dazu müssen eventuell ungenaue Beschreibungen oder Bezeichnungen der Nutzenden auf die vorhandenen Bezeichnungen von Haltestellen abgebildet werden, was wieder – wie bei einem IR-System – die Verarbeitung vager Angaben notwendig machen kann.

1.6 Management-Informationssysteme

Salton und McGill (1983) [103] führen als weitere Beispiele von Informationssystemen Management-Informationssysteme (*management information systems*) und Entscheidungsunterstützungssysteme (*decision support systems*) auf. Dabei handelt es sich weniger um eine Beschreibung des zugrunde liegenden Modells, sondern eher um eine Charakterisierung durch die Inhalte, die das System anbietet, und die Art und Weise ihrer Präsentation. Sie sind spezifisch auf den Einsatz in großen Organisationen und den Informationsbedarf in deren Management ausgerichtet. Entscheidungsunterstützungssysteme sind darüber hinaus insbesondere dazu geeignet, verschiedene Handlungsalternativen zu unterscheiden und Prognosen über zu erwartende Entwicklungen zu bieten.

Begriffe wie *Data Warehouse* und *Intranet* beschreiben den kontrollierten und vereinheitlichten Zugriff auf Dokumente und Informationen in einer Organisation, die z. B. in verschiedenen relationalen Datenbanken, Dokument-Servern oder spezialisierten Informationssystemen gespeichert sind.

Unter der Bezeichnung *OLAP* (*On-Line Analytical Processing*) werden Systeme zusammengefasst, mit denen Datenbankinhalte bzw. bestimmte Ausschnitte oder Projektionen der Datenbankinhalte grafisch dargestellt werden können. Mit OLAP-Systemen lassen sich häufig auch die Auswirkungen von Änderungen bestimmter Parameter wie Preise oder Verkaufszahlen sichtbar machen. Solche Visualisierungswerkzeuge können auch der Entscheidungsunterstützung dienen. Auf sie wird hier allerdings nicht weiter eingegangen.

Expertensysteme und Entscheidungsunterstützungssysteme beschränken sich nicht mehr darauf, einzelne in sich geschlossene Informationsobjekte auszuwählen und anzuzeigen. Bei diesen Systemen werden die zugrunde liegenden Daten zur Anfragezeit aus einer Datenbasis erzeugt und in die Form gebracht, in der sie präsentiert werden. Bei Expertensystemen müssen die Fakten dazu in spezifischen Formaten vorliegen und gegebenenfalls in sich konsistent sein. Ähnliches gilt bei Entscheidungsunterstützungssystemen, sobald Prognosen über zu erwartende Entwicklungen getroffen werden müssen. Hier werden mit vorhandenen Daten in einem spezifischen Format vorher festgelegte Extrapolationen berechnet.

1.7 Data Mining

Ein in gewisser Weise umgekehrtes Szenario wird bei der *Wissensgewinnung aus Datensammlungen* (*Knowledge Discovery in Databases* oder *Data Mining*) verwendet. Hier wird in vorhandenen Datensammlungen nach nützlichen Regelmäßigkeiten gesucht. Dabei braucht zunächst nicht bekannt zu sein, welche Eigenschaften oder Attribute der Datensätze wichtig sind und welche nicht. Die Data-Mining-Verfahren sollen gerade das herausfinden. Allgemein lässt sich das so formulieren: *Knowledge Discovery in Databases (*KDD*) beschreibt automatisierte Verfahren, mit denen Regelmäßigkeiten in Mengen von Datensätzen gefunden und in eine für Nutzende verständliche Form gebracht werden.*

Der erste Teil dieser Definition ist auch eine Beschreibung von *Machine Learning* (ML), einem Forschungsgebiet, das sich damit befasst, Verfahren zu entwickeln, die mit einer Menge von Beispielen *trainiert* werden und anschließend in der Lage sind, diese und andere Beispiele (möglichst) richtig zu bearbeiten. Eine klassische Aufgabe besteht darin, Objekte, die durch einen Datensatz beschrieben sind, verschiedenen Klassen zuzuordnen, sie also zu klassifizieren. Einige Autoren definieren Knowledge Discovery in Databases als Machine Learning, bei dem die Trainingsmenge eine Datenbank ist (Holsheimer und Siebes, 1994 [56]). Dabei bleibt allerdings die zweite Forderung – dass nämlich die gefundenen Regelmäßigkeiten für Menschen verständlich sein müssen – unberücksichtigt. Verständliche Ergebnisse können z. B. die Form haben: »Wenn X und Y, dann Z« oder »in $X\,\%$ der Fälle, in denen Y eintritt, tritt auch Z ein«.

Das Klassifizieren von Objekten – in diesem Fall Dokumenten – ist auch eine zentrale Aufgabe des Information Retrieval, wenn es z. B. darum geht, Artikel oder Bücher bestimmten Themenklassen zuzuordnen. Wichtige Teilaufgaben aus den Gebieten Machine Learning und Information Retrieval lassen sich formal also mit demselben Modell beschreiben. Wie weit sie auch mit denselben Verfahren bewältigt werden können, wird in diesem Buch untersucht.

Knowledge-Discovery-Systeme arbeiten häufig auf Faktendatenbanken, also mit stark strukturierten Datensätzen. Der Typ der Einträge in diesen

Datensätzen ist meist wohl definiert, also beispielsweise ein binärer, ganzzahliger oder reeller Wert wie Verkaufsdaten, Umsatzzahlen, Testwerte oder Preise. Als Ergebnisse liefern sie z. B. Verfahren, mit denen neue Datensätze in Kategorien eingeteilt werden können. So kann man z. B. versuchen, die Kundendaten einer Kreditkartengesellschaft zu verwenden, um bessere und einheitlichere Regeln für die Aufnahme neuer Kunden zu ermitteln (Carter und Catlett, 1987 [22]).

1.8 Kategorisierung mit einem Data-Mining-System

Carter und Catlett (1987) [22] beschreiben ein Machine-Learning-Programm, das Entscheidungen darüber trifft, ob ein Antrag auf eine Kreditkarte bewilligt werden soll oder nicht. Traditionell können diese Entscheidungen von den Mitarbeitenden des jeweiligen Instituts aufgrund ihrer Erfahrungen und anhand von Richtlinien getroffen werden; oder es werden *Bewertungstabellen*, so genannte *Scoring Tables* (siehe Abbildung 1.4), verwendet.

	boarder	rent	mortgage	owner
Home status	5	8	15	20
Time at address	0 - 1 *4*	1 - 2 *7*	2 - 3 *10*	3 - 4 *15*
Age of car	None *0*	0 - 1 *10*	1 - 2 *15*	2 - 3 *11*
Monthly disposable income	0 - $124 *0*	$125 - 249 *10*	$250 - 349 *15*	$350 up *25*

Abbildung 1.4 – Scoring Table: In der linken Spalte stehen die Attribute, in den Zellen jeweils oben die möglichen Werte und unten (kursiv) die dazugehörigen Punktzahlen. »Home status« beschreibt die Wohnsituation. Ein »boarder« ist z. B. ein Untermieter, »mortgage« ist eine Hypothek; dieser Attributwert beschreibt also vermutlich jemanden, der eine eigene hypothekenbelastete Wohnung bewohnt. (Aus Carter und Catlett, 1987 [22])

Mit diesen Tabellen werden für bestimmte Attribute – wie Höhe des Einkommens, Höhe des Bankguthabens oder Grundbesitz – Punkte vergeben. Wenn die Summe der Punkte einen Schwellwert übersteigt, wird eine Kreditkarte vergeben, sonst nicht. Anstelle dieses einfachen Summationsverfahrens kann auf der Basis von Fallbeispielen mit Machine-Learning-Verfahren ein Entscheidungsbaum erzeugt werden (siehe Abbildung 1.6).

Carter und Catlett machen leider keine genaueren Angaben über die verwendeten Attribute (vermutlich, weil das System tatsächlich angewendet wur-

No	Attributes				Class
	account	balance	employed	monthly expense	
1	bank	700	yes	200	accept
2	bank	300	yes	600	reject
3	none	0	yes	400	reject
4	other inst	1200	yes	600	accept
5	other inst	800	yes	600	reject
6	other inst	1600	yes	200	accept
7	bank	3000	no	300	accept
8	none	0	no	200	reject

Abbildung 1.5 – Eine kleine Trainingsmenge: In jeder Zeile stehen die Attributwerte für einen Antrag; in der letzten Spalte rechts steht, ob der Antrag akzeptiert wurde (accept) oder nicht (reject). (Aus Carter und Catlett, 1987 [22])

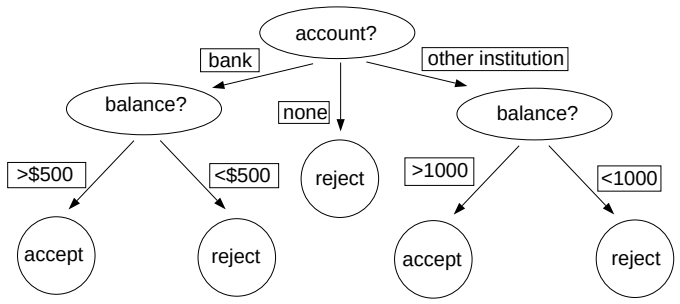

Abbildung 1.6 – Entscheidungsbaum zur Trainingsmenge aus Abbildung 1.5: Um ein neues Beispiel zu kategorisieren, also eine Entscheidung zu treffen, wird (von oben beginnend) das Attribut geprüft, das im jeweiligen Knoten (als Ellipse dargestellt) steht. Für jeden möglichen Wert dieses Attributs gibt es einen Kindknoten, zu dem übergegangen wird, falls das Attribut im Beispiel diesen Wert hat. Wird ein Endknoten (Kreis) erreicht, liegt eine Entscheidung vor. (Nach Carter und Catlett, 1987 [22])

de), sondern geben nur hypothetische Beispiele. So ist in Abbildung 1.4 eine Scoring-Tabelle dargestellt und in Abbildung 1.5 eine Trainingsmenge. Abbildung 1.6 zeigt einen Entscheidungsbaum, der die Beispiele aus der Trainingsmenge aus Abbildung 1.5 richtig kategorisiert. Man beachte, dass (in diesem hypothetischen Beispiel) nur zwei der vier Attribute verwendet werden.

Andere Beispiele für Regeln sind Regelmäßigkeiten im Kaufverhalten von Konsumenten, oder die Analyse von Verbindungsdaten von Funktelefonen, um den Missbrauch von Kennungen festzustellen (siehe Kapitel 8). Häufig werden auch (anonymisierte) Patientendaten verwendet, um ärztliche Diagnosen aus den verschiedenen Labor- und Testwerten vorherzusagen.

Diese Beispiele zeigen auch die Problematik des Ansatzes. Es gehört nicht viel Phantasie dazu sich vorzustellen, wie die Verfahren z. B. bei der Vergabe von Arbeitsplätzen, dem Abschluss von Versicherungsverträgen oder der polizeilichen (Raster-)Fahndung angewendet werden können. Dabei besteht natürlich die Gefahr, dass aufgrund allgemeiner Regeln im Einzelfall falsche Schlüsse auf die individuellen Fähigkeiten, Verhaltensweisen, Eigenschaften und Risiken einer Person gezogen werden.

Aber auch wenn die Regeln und Schlüsse richtig sein sollten, wird sich eine Gesellschaft grundlegend verändern, wenn ihre Bürgerinnen und Bürger durch die Anwendung der Regeln auf allgemein zugängliche oder leicht zu erhebende Daten sehr genau eingeordnet werden können.

Schließlich kann allein der Glaube von Entscheidungsträgern an die Vorhersagen eines Systems bei Ermessensentscheidungen ausschlaggebend sein und zur Diskriminierung ganzer Gruppen führen.

1.9 Assoziative Regeln und der Warenkorb

Andere häufig untersuchte Regelmäßigkeiten sind *assoziative Regeln*, die z. B. aus Verkaufsdaten gewonnen werden können. Eine Datensammlung, aus der solche Regeln gewonnen werden, besteht aus Datensätzen, die Teilmengen einer Grundmenge beschreiben, also z. B. einzelne Einkäufe aus dem Sortiment eines Ladens oder eines Versandhauses. Ziel ist es, typische *Warenkörbe* zu bestimmen, also Gruppen von Artikeln, die häufig zusammen gekauft werden, bzw. für eine Menge von Waren zu bestimmen, welche weiteren Waren typischerweise gekauft werden.

Dazu wird zum einen für eine Teilmenge der Grundmenge untersucht, in wie vielen der Datensätze sie auftritt; zum anderen wird festgestellt, wie sich diese Zahl verändert, wenn ein Artikel weggelassen wird. Ist die Anzahl des Auftretens groß und steigt sie nur wenig an, wenn ein Artikel weggelassen wird, ergibt sich eine assoziative Regel.

Würden z. B. in 400 von 1 000 Einkäufen Eier, Salz, Butter, Schmalz, Milch, Mehl und Safran gekauft und in 500 Einkäufen alle diese Artikel außer Safran, so hätte die assoziative Regel

```
Eier, Salz, Butter, Schmalz, Milch, Mehl -> Safran
```

eine Basis von $0,4$ und eine Sicherheit von $0,8$ und würde auf der Ebene des Einkaufs ein altes Kinderlied neu entdecken.

Welcher Nutzen für den Verkauf aus einer solchen Regel gezogen werden kann, bleibt zunächst offen. Bei der Anwendungen von assoziativen Regeln im Information Retrieval können so z. B. neue Suchterme gefunden und vorgeschlagen werden (siehe Kapitel 7).

1.10 Wissensgewinnung und Information Retrieval

Klassische IR-Systeme bedienen einen durch eine Anfrage ausgedrückten Informationsbedarf mit Dokumenten oder Datensätzen aus einer Sammlung oder Datenbank, die mehr oder weniger den tatsächlichen Informationsbedarf der Anfragenden befriedigen. Dabei wird aber in der Regel das Verhältnis der Dokumente oder Datensätze untereinander nicht weiter berücksichtigt. Um Regelmäßigkeiten zwischen den Dokumenten oder Datensätzen zu nutzen, können Wissensgewinnungsverfahren eingesetzt werden. Dabei stehen dann nicht mehr die einzelnen Dokumente oder Datensätze im Vordergrund, sondern die Daten werden quasi als Rohstoff verwendet, um daraus neues Wissen zu gewinnen, mit dem die Suche nach Dokumenten unterstützt werden kann. Als Beispiel wurden bereits assoziative Regeln erwähnt, mit denen weitere Suchwörter zu einem Thema gefunden werden können. Es gibt aber noch weitere Möglichkeiten, Data-Mining-Methoden für das Information Retrieval zu nutzen.

Anfrage	IRS als Indexterm	Anzahl
`Information retrieval system`		92
`Information retrieval systems`		219
	ja	203
`Information retrieval system` `AND Information retrieval systems`		30
`Information retrieval system`	nein	68
`Information retrieval systems`	nein	16
`Information retrieval system`	ja	24

Abbildung 1.7 – Anzahl der in INSPEC gefundenen Dokumente für das erste Halbjahr 1995: Die zweite Spalte (IRS als Indexterm) gibt an, ob der Indexterm »information retrieval systems« nicht beachtet wird, vorhanden sein muss (ja) oder nicht vorhanden sein darf (nein).

Einige einfache Beobachtungen kann man schon aus Anfragen an eine herkömmliche Literaturdatenbank ableiten: Zum Beispiel findet INSPEC (Januar bis Juni 95) für die Anfrage `Information retrieval system` 92 Einträge, für die Anfrage `Information retrieval systems` (also den Plural) 219, für die Schnittmenge der beiden Anfragen aber nur 30. Die Singular- und die Pluralformen des Suchterms `Information retrieval system` haben also sehr verschiedene Treffermengen in INSPEC.

Information retrieval systems ist ein Indexterm von INSPEC. Er sollte als solcher den Dokumenten zugewiesen worden sein, die sich hauptsächlich mit diesem Thema beschäftigen und bei denen es nicht nur am Rande erwähnt wird. Weitere Trefferzahlen für Anfragen, bei denen die Benutzung als Index-

term einbezogen wurde, finden sich in Abbildung 1.7. Es zeigt sich, dass die relative Häufigkeit des Indexterms bei der Pluralform ca. zehnmal so hoch ist wie bei der Singularform. Über Gründe für dieses Einzelergebnis kann man natürlich nur spekulieren. Vielleicht werden von dem System, mit dem die Artikel indexiert werden, im Text (Titel oder Abstract) auftretende Indexterme automatisch erkannt und den (menschlichen) Indexierenden vorgeschlagen. Vielleicht ist es aber auch so, dass die Pluralform eher verwendet wird, wenn über ein System berichtet wird, und die Singularform, wenn es nur als Beispiel am Rande erwähnt wird. Auch hier lässt sich ein Problem von KDD-Verfahren beobachten: Wenn ein Zusammenhang gefunden und beschrieben wird, müssen die Gründe für sein Auftreten noch nicht erkannt sein. Sie können in Umständen liegen, die zwar systematisch bei den Einträgen einer Datensammlung auftreten, aber deshalb noch lange nicht charakteristisch für die beschriebenen Objekte oder Zustände sein müssen.

Data Mining und Information Retrieval haben sich zunächst eher unabhängig voneinander entwickelt. Der Zusammenhang zwischen den beiden Forschungsgebieten wird allerdings zunehmend wahrgenommen und genutzt. Bei der Kombination von Methoden aus beiden Gebieten spricht man auch von *Text-Mining*-Verfahren.