Dr. Felix Engel

Modul 63415

Information Retrieval

LESEPROBE

Fakultät für
Mathematik und Informatik

FernUniversität in Hagen

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Vorwort

Dieses Skript ist in Beratung mit Prof. Dr.-Ing Hemmje entstanden. Für dessen wertvolle Hinweise ich sehr dankbar bin. Weiterhin möchte ich mich auch bei unserem Doktoranden Christian Nawroth, für seine ausführlichen Anmerkungen bedanken.

Textuelle Beiträge wurden auch von studentischen Arbeiten entnommen die von mir in Ihrer Abschlussarbeit betreut wurden. Ich danke auch den Studenten an dieser Stelle sehr herzlich Kurseinheiten mit textuellen Beiträgen von Studenten werden an entsprechender Stelle ausgezeichnet.

Aufgrund der festen Verankerung einer bestimmten Terminologie in der IR Gemeinschaft werden in diesem Kurs einige engl. Bezeichnung einer deutschen Übersetzung vorgezogen. Zunächst wird jedoch immer auch eine deutsche Übersetzung eingeführt, im weiterem Verlauf aber auf die englische Übersetzung zurückgegriffen.

1 Einführung in das Information Retrieval

Beständig- und im zunehmend Maße werden Informationen direkt in digitaler Form erzeugt oder nachträglich in ein digitales Format überführt. Ein Grund dafür ist die schnelle und einfache Verarbeitung und eine damit einhergehende bessere Wiederverwendbarkeit. Einen umfangreichen digitalen Datenbestand jedoch manuell und gezielt nach einer bestimmten Information zu durchsuchen ist ab einer bestimmten Menge an Daten nicht mehr effektiv möglich und der tatsächliche Nutzen des Bestands damit zumindest fraglich. Ein plakatives Beispiel für einen sehr umfangreichen und multimedialen Datenbestand ist das Internet, welches massive Mengen an digitalen Daten vorhält. Wohlbekannte Suchmaschinen helfen hier dem suchenden, um sich in diesem Bestand zurechtzufinden. Große Datenbestände entstehen jedoch auch in spezielleren Bereichen, wie z.B. in Behörden, Krankenhäusern oder Verlagen. Auch hier muss ein effektives Auffinden gesuchter Informationen gewährleistet werden. Die Lehre im Umfeld des *Information Retrieval* (IR) befasst sich daher mit der Modellierung und Umsetzung von Anwendungen die automatisiert digitalen Datenbestände, für den einfachen Zugriff und Nachnutzung aufbereiten. Die Lehre von effektiven IR Prozesse ist hinreichend komplex und obwohl das IR auf eine lange Historie zurückblick sind insbesondere mit Hinblick auf anwachsende Datenmengen, mit zunehmend heterogener und verteilter Natur, Fragestellungen offengeblieben und neue Anforderungen hinzugekommen.

1.1 Aufbau, Einordnung des Inhalts und Lernziele

1.1.1 Aufbau

Die Kurseinheit 1 umfasst fünf inhaltliche Kapitel und deren zugehörige Unterkapitel. Zum Ende der Kurseinheit wird es zu jedem der Kapitel Verständnisfragen zu den behandelten Themenbereichen geben.

1.1.2 Einordnung

Die erste Kurseinheit befasst sich mit der Darstellung der generellen Motivation für das IR, den Grundlagen die den Schwerpunkt der Themen aller weiteren Kurseinheiten bilden und einer Einordnung in den akademischen Kontext. Inhaltlich führt die erste Kurseinheit dann auch in Anforderungen an des IR ein, welche den grundlegenden Rahmen an das IR selbst abstecken und eine generelle Sichtweise auf Lösungen des IR geben. Zunächst wird daher diskutiert was „Information" im generellen und Information Retrieval im speziellen bedeutet. Aufbauend auf diesen grundlegenden Erläuterungen folgt eine erste Sicht auf ein konzeptuelles IR Modell, welches prinzipiell das Vorgehen der Gruppen an IR Prozesse zeigt welche Bestandteil dieses Kurses sind.

Die drei letzten Kapitel dieser Einheit beschreiben weitere Grundlegende Bestandteile des IR. Dies ist zum einen das Konzept des Informationsbedarfs, welches grundlegenden Einfluss auf die Initiierung einer Suche hat und damit starken Einflussnahme auf das IR selbst nimmt. Wird die Initiierung der Suche betrachtet muss auch erläutert werden was überhaupt über welche Mechanismen gesucht werden kann. Diese beiden Themen (was kann gesucht werden und über welche generellen Mechanismen) werden abschließend in den letzten Kapiteln der Kurseinheit besprochen. Die erste Kurseinheit ist somit eher theoretischer Natur. Die weiteren Kapitel beschreiben dann zunehmend praktischere Anwendungen des Themenbereichs.

1.1.3 Lernziele

Das übergeordnete Lernziel des Kurses 1879 ist es ein grundlegendes Verständnis für die vielfältigen Themenfelder, welche das IR ausmachen, zu erreichen. Dazu gehört ein sicherer Umgang mit der Fachterminologie und gutes Verständnis der Basiskonzepte und grundlegenden Modellen. Darüber hinaus soll ein sicherer Umgang bei der Bewertung von IR Prozesse erreicht werden. Letztlich soll ein gutes Verständnis über die die klassischen IR Verfahren erreicht werden und darüber hinaus ein Einstieg in die Semantische Suche erfolgen. Das Vorgestellte soll dabei konzeptuell und praktisch erlernt werden.

Lernziel der vorliegenden Einheit ist es zunächst einen sicheren Umgang mit relevanter Terminologie und grundlegender Konzepte zu erlangen.

1.2 Information Retrieval

Im akademischen Umfeld ordnet sich die IR Lehre in den Fachbereich der Informatik ein und wird im thematischem Umfeld der *Informationssysteme* (IS) behandelt. Ein Blick auf das etablierte *Computing Classification System* (siehe [ACM19]) offenbart dabei die weitere Untergliederung von IS zu IR und in weitere verwandte Felder, wie z.B. *Data Management Systems* oder *Information Storage Systems*. Die Lehre im Umfeld des IR ist thematisch sehr breit aufgestellt und umfasst eine große Menge unterschiedliche Bereiche. Das IR ist im akademischem Umfeld keine neue Wissenschaft, sondern besteht schon seit ca. Mitte des 20. Jahrhunderts. Über die Jahrzehnte hat sich dabei eine feste Terminologie eingebürgert, die es erleichtert sich über Inhalte des IR auszutauschen. Die nächsten Kapitel erläutern zunächst was genau unter „Information Retrieval" zusammengefasst wird und benennt einige Anwendungen des IR.

1.2.1 Was ist Information?

Der Begriff der *Information* ist zentral in der Einordnung und für das Verständnis des IR. Wie in vielen wissenschaftlichen Disziplinen gibt es auch hier eine Menge an verschiedenen Definitionen. Eine allg. Definition wird zunächst in gängigen Lexika gefunden. U.a. in Merriam Webster: "*the communication or reception of knowledge or intelligence*". Information wird damit als das Empfangen von Wissen umschrieben.

Auf dieser Basis ist es insofern wichtig direkt eine Abgrenzung zu dem eng verwandten Begriff des *Wissens* und auch von digitalisierten *Daten* als Grundlage des IR an sich vorzunehmen. U.a. der Duden definiert *Daten* als: „*(durch Beobachtungen, Messungen, statistische Erhebungen u. a. gewonnene) [Zahlen]werte, (auf Beobachtungen, Messungen, statistischen Erhebungen u. a. beruhende) Angaben, formulierbare Befunde*" [Dd]. *Wissen* hingegen ist ein wesentlich komplexeres Konzept, dass eine enge Verbindung zu weiteren Konzepten, wie U.a. der *Fähigkeit* und *Kompetenz* eines Menschen aufweist. Als eine Vertreterin der Wirtschaftswissenschaften, vertritt Rowley in der vielzitierten Publikation [R07], die Anschauung das Wissen auf Basis der Definition von Daten und Information abgeleitet wird. Wissen wird dabei unter anderem als eine Kombination aus Daten, Informationen und persönlichem Kontext verstanden und wird dazu genutzt Entscheidungen zu treffen. In diesem Umfeld ist Konsens, dass die folgende Verkettung gilt:

Daten → Information → Wissen → Entscheidung

Eine weitere interessante Definition stammt von Nauta [N19], welcher den subjektiven und temporalen Aspekt hervorhebt der mit der Definition von Information einhergeht: "*Information is news: what is known already is no information to the extent that it is unknown, unexpected, surprising, or improbable*". Nauta wirft in diesem Zusammenhang auch weitere interessante Fragen auf, welcher wir aber in dieser Kurseinheit nicht weiter behandeln werden, aber zur weiteren Diskussion zumindest benennen: "*can there be information without language?*" or "*what does genetic information have in common with newspaper information?*".

In diesem Skript wird jedoch die Definition von Kuhlen, aus dem Blickwinkel der Informationswissenschaften [HK90] Verwendung finden. Kuhlen stellt eine Definition aus semiotischer (Wissenschaft der Zeichensysteme, u.a. Sprache) Sicht auf. Dies, weil im IR im speziellem Dokumenten betrachtet werden, deren Inhalt Sprache ist die in Form von Text repräsentiert wird. Er definiert Information als *„...die Teilmenge von Wissen, die von einer bestimmten Person oder einer Gruppe in einer konkreten Situation zur Lösung von Problemen benötigt wird und häufig nicht vorhanden ist."*. Kuhlen unterscheidet *Wissen* als den Zustand einer Person, der unabhängig von einem Kommunikationspartner ist. *Information* fasst er im Gegensatz als einen Zustand der Kommunikation zwischen Personen auf. Kuhlen stellt damit fest, dass die Information eine Teilmenge von Wissen ist. Aus seiner Sicht gilt: Daten sind die Basis um Wissen zu erzeugen und aus Wissen können wieder Daten entstehen. Informationen hingegen leiten sich wiederum aus Wissen ab, können aber auch wieder zu Wissen werden (vergl. Abbildung 1-1).

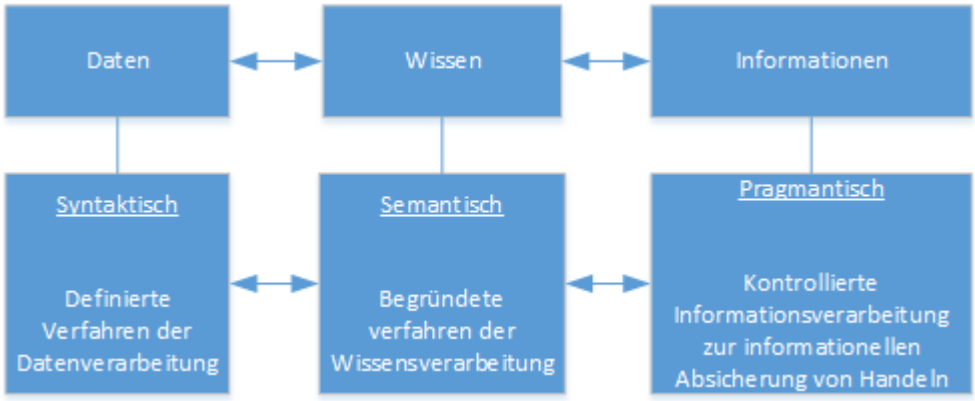

Abbildung 1-1: Zusammenhang zwischen Daten, Wissen und Informationen [K85]

Kuhlen definiert, dass Daten rein auf syntaktischer Ebene angesiedelt sind. Daten sind als solches über Regelwerke beschreibbar. Wissen jedoch umfasst über die Definition von Daten hinaus semantische Zusammenhänge. Informationen, zu guter Letzt sind aus Wissen abgeleitet und handlungsrelevant. Für den weiteren Verlauf dieses Kurses wird diese Sichtweise von Kuhlen eingenommen.

1.2.2 Was ist Information Retrieval

Nachdem der Begriff der Information eingeführt und in Kontext gesetzt wurde, stellt sich weiterhin die Frage was sich hinter der Zusammensetzung der Wörter *Information* und *Retrieval* verbirgt. Um IR in Abgrenzung zu anderen Themen der Informatik einzuführen ist es dabei interessant erst einmal eine direkte deutsche Übersetzung des englischen Begriffes „Information Retrieval" zu betrachten. Lexika belegen diesen Begriff mit unterschiedlichen Beschreibungen, wie *„die Informationsgewinnung"*, *„Wiederauffinden von Informationen"* oder auch *„Informationswiedergewinnung"*. Naheliegend ist zumindest, dass der Begriff der Information und dessen Verarbeitung eine Maßgebliche Rolle spielen. Zunächst sei eine bekannte Definition von Baeza-Yates gegeben. Er definiert in [BYRN99]: *"(IR) part of computer science which studies the retrieval of information (not data) from a collection of written documents. The retrieved docu-*

ments aim at satisfying a user information need usually expressed in natural language". Das IR beschreibt damit generell gesprochen einen Prozess der automatisierten Unterstützung bei der Suche nach Informationen leistet. Ein Prozess bezeichnet hierbei eine zunächst ganz allgemein eine geordnete Menge von Verarbeitungsschritten, die sukzessive abgearbeitet werden. Im IR wird der Prozess wird durch einen *Anwender* (engl. User) initiiert, weil dieser einen Bedarf nach Informationen hat. Etwas konkreter, aber sehr ähnlich der Definition von Baeza-Yates definieren Manning et al. in [MRS08] die allgemeine Anforderung an das IR als: *"Information Retrieval (IR) is finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers)*". Das folgende Kapitel wird zunächst noch einmal darauf eingehen, welche Bedeutung dem Informationsbedarf im IR zukommt.

1.3 Informationsbedarf und Information Seeking

Dieses Kapitel erläutert was der Informationsbedarf für das IR bedeutet und welche Wichtigkeit er für die automatisierte Suche nach Dokumente hat. Das Kapitel 1.2.2 hat bereits zwei bekannte IR Definitionen eingeführt. Aus beiden Definitionen wird deutlich, dass der Informationsbedarf eines Users Auslöser für die Nutzung eines IR Prozesses ist: „*finding material … that satisfies an information need*".

Manning et al. beschreiben in [MRS08] den Informationsbedarf folgend: "*An information need is the topic about which the user desires to know more, and is differentiated from a query, which is what the user conveys to the computer in an attempt to communicate the information need*". Der *Informationsbedarf*, der über ein IR Prozesse aufgelöst werden soll, entspricht demnach dem Gegenstand der Betrachtung, über welchen der User eines IR-Prozesses in Unsicherheit ist, bzw. sein Wissen erweitern möchte. Aus der angeführten Beschreibung geht weiterhin hervor, dass der Informationsbedarf in Form einer *Anfrage* (engl. Query) formuliert und an den IR-Prozess übergeben wird. Der Informationsbedarf und die Query müssen jedoch als zwei separate Eigenschaften betrachtet werden. Denn die Query ist nur das Transportmittel um dem Informationsbedarf Ausdruck zu verleihen und es muss ein Übersetzungsprozess vom Informationsbedarf in die Query stattfinden.

Die Übersetzung bzw. die Formulierung der Query unterliegt dabei immer einer gewissen Ungewissheit oder Vagheit. Insbesondere, da es potentiell viele Möglichkeiten gibt sich auszudrücken oder es gar nicht einfach möglich ist, den Bedarf in Worte zu fassen. Vor allem wenn er vom Menschen an eine Maschine kommuniziert (übersetzt) werden muss. Ein bekanntes, grundlegendes Modell, welches den Informationsbedarf und die Interaktion eines Menschen mit einem Informationssystem beschreibt ist das sog. *Anomalous States of Knowledge* Modell von Nicholas J. Belkin (siehe [B80]) das folgend eingeführt werden soll.

Anomalous States of Knowledge (ASK) ist ein Problem orientiertes Modell und basiert auf der Annahme das ein menschlicher Erzeuger von Daten und ein menschlicher User dieser Daten,

effizient miteinander Informationen austauschen möchten. Diese Annahme schränkt damit bewusst das Problemfeld auf die menschliche Kommunikation folgend ein:

1) Der User eines IR Prozesses erkennt einen Informationsbedarf und übergibt eine entsprechende Suchanfrage an das System, um seinen Bedarf zu decken.
2) Die Aufgabe des IR-Prozesses ist es dem User ein Dokument mit textuellem Inhalt (als ein klassisches Beispiel eines Informationsträgers) zu liefern, das höchstwahrscheinlich den Bedarf des Users entspricht.
3) Der User untersucht den gelieferten Text, womit der Informationsbedarf zu einem bestimmten Grad gedeckt wird oder auch gar nicht adressiert werden konnte. Die Beurteilung des Users wird anhand der Relevanz bemessen, welche die gefundenen Texte zur Suchanfrage haben.

Das Kommunikationsmodell das ASK beschreibt, ist schematisch in Abbildung 1-2 dargestellt. Auf der linken Seite befindet sich hierbei der Erzeuger, welcher sein Wissensstand und Sicht auf die Welt in Form von als Text manifestiert hat. Auf der rechten Seite hingegen befindet sich der User, mit einem abweichenden Wissenszustand, welcher sich in einer Frage (Request) manifestiert.

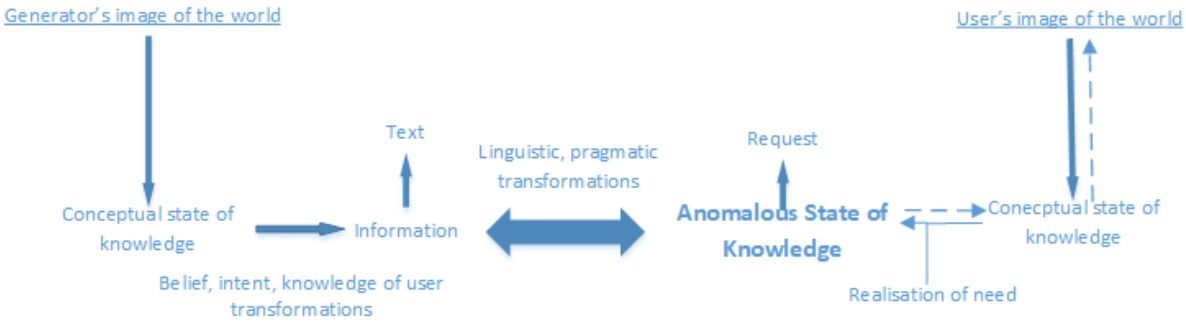

Abbildung 1-2: Ein kognitives Kommunikationsmodell nach Belkin [B80]

Das ASK Modell beschreibt hinlänglich, dass sich die inhärente Schwierigkeit die ein IR-Prozess zu bewältigen hat aus unterschiedlichen Wissensständen der beteiligten Kommunikationspartner ergibt. Belkin beschreibt daher, dass ein wichtiger Aspekt in der Erklärung des Informationsbedarfs, dass Erkennen unzulänglichen Wissens eines Users ist. Insbesondere erklärt er damit auch das die iterative Interaktion eines Users im IR Prozess wesentlich Bestanteil ist.

Der Informationsbedarf und die Formulierung der Suchanfrage werden hier als zwei getrennte Konzepte betrachtet, denn die Suchanfrage versucht „lediglich" den Bedarf eines Users abzubilden. Modelle, welche der iterativen Interaktion des Users mit dem IR-Prozess beschreiben werden thematisch dem sog. *Information Seeking* zugeordnet. Die folgenden Texte geben einen Überblick über Erkenntnisse aus dem Umfeld des Information Seeking. Ellis Kuhltau fand, bei einer Untersuchung des Verhaltens bei der Informationssuche von wissenschaftlichen Mitarbei-

tern heraus, dass ein Muster aus sechs wiederkehrende Merkmalen unterschieden werden können (vergl. [E89]). Die sechs Merkmale im einzelnem können der Tabelle 1-1 entnommen werden. In welcher Verbindung die einzelnen Merkmale zueinanderstehen, wurde offengelassen. Es wurde lediglich angemerkt, dass die Ausprägung und Verbindungen Kontextabhängig von Person und Art der Suche sind.

Merkmale	***Systemdesign***
Start (engl. Starting): Aktivitäten die der Initiierung der Suche zugeordnet werden können.	Wird ein neues Projekt gestartet sollen relevanten Referenzen zum Thema verfügbar sein, die z.B. Schlüsselstudien referenzieren, einen Überblick verschaffen oder als Basis einer Verkettung dienen. Falls derlei Informationen nicht initial vorhanden sind, sollten diese über das IR verfügbar gemacht werden.
Verketten (engl. Chaining): Folgen von Referenzen (z.B. Fußnoten oder Literaturverzeichnis)	Das einfache Verfolgen von Referenzen soll unterstützt werden.
Durchstöbern (engl. Browsing): vage Suche in im Bereich des potentiellen Interesses	Unterstützung von semistrukturierter Suche z.B. über Autorennamen, Konferenzen oder Schlagworte.
Abgrenzung (engl. Differentiation): Ausnutzen von Unterschieden zwischen dem bereits untersuchten Material um einen Überblick über die Eigenschaft und Qualität des Materials zu bekommen.	Werden unterschiedliche Quellen in die Suche aufgenommen kann eine Einteilung in relevante und weniger relevante Quellen hilfreich sein, um die Mange an verfügbaren Material zu kondensieren.
Überwachung (engl. Monitoring): Durch Überwachen einer bestimmten Quelle, über alle Neuerungen informiert bleiben.	Das System sollte eine Liste an Quellen und Änderungen bereitstellen.
Extrahieren (engl. Extracting): Systematische Analyse einer Quelle nach relevantem Material	

Tabelle 1-1: Information Seeking Merkmale von E. Kuhltau

Das Model von M. Bates (vergl. [B02]) stellt den Menschen und seinen Werdegang in das Zentrum des Information Seeking Prozesses und betrachtet diesen Vorgang unter Berücksichtigung von Erkenntnissen aus der Kunde des Menschen (Anthropologie). „*Looking at us as a species that exists physically, biologically, socially, emotionally, and spiritually, it is not unreasonable to*

guess that we absorb perhaps 80 percent of all our knowledge through simply being aware, being conscious and sentient in our social context and physical environment.". Sie erkennt verschiedene Modi dem Einfluss auf das Verhalten bei der Suche nehmen. Ein Kernelement Ihres Modells dabei ist das Informationen nicht nur aktiv, sondern auch passiv über die Zeit aufgenommen werden.

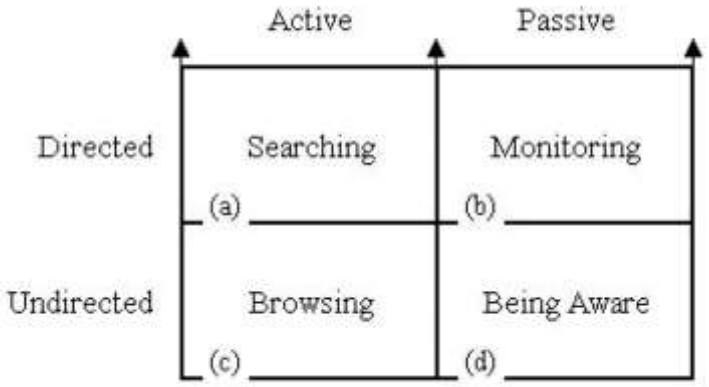

Abbildung 1-3: Bates Information Seeking and Searching Model

- **Bewusstsein (engl. Awareness):** das was wir aufnehmen und lernen über passives Verhalten.
- **Überwachen (engl. Monitoring):** Überwachen und Stöbern sind sich ergänzende Modi. Überwachen ist direkt und passiv, während Stöbern ungerichtet und aktiv betrieben wird.
- **Stöbern (engl. Browsing):** Es gibt keinen spezifischen Informationsbedarf, sondern es findet eine aktive Suche nach neuen Informationen statt.
- **Suchen (engl. Searching):** ein aktives Bestreben Antworten zu offenen Fragen zu finden oder ein Verständnis für einen neuen Sachverhalt zubekommen.

Zuletzt soll noch eine aktuellere, vielzitierte Studie von Hemminger et al. (vergl. [HLVA07]) dokumentiert werden. Diese zeigt, ergänzend zu den vorab vorgestellten Modellen, welchen Einfluss die allgegenwärtige Verfügbarkeit Internetbasierter Angebot auf das Information Seeking Verhalten von akademischen Wissenschaftlern hat. Das Ergebnis zeigte klar, dass mit zunehmenden Angebot dieser Dienste, diese auch zunehmend in Anspruch genommen werden. In Konsequenz führte dies zu weniger Besuchen einer Bibliothek und fast vollständiger Kommunikation über das Internet. Diese Entwicklung wird u.a. durch die in Abbildung 1-4 dargestellte Tabelle ersichtlich. Abbildung 1-5 zeigt hingegen eine Statistik zur Verwendung spezifischer Werkzeuge. Die angedeutete Situation spiegelt sich nicht zuletzt seit geraumer Zeit im wissenschaftliche Publikationswesen wieder. Fast ausschließlich werden wissenschaftliche Erkenntnisse über elektronische Medien verbreitet, eher selten als gedruckte Auflage.

Das Information Seeking hat insofern einen großen Einfluss auf das IR, insbesondere im interaktivem IR (vergl. [UG19]) und setzt das Verhalten des menschlichen Users eines IR Prozesses in den Vordergrund. Dieses Skript wird im weiterem Verlauf nur punktuell auf das Information Seeking zurückgreifen, da der Fokus eher auf Algorithmen und Datenstrukturen liegt, welche den Kern im IR ausmachen.

	Daily or Weekly %	Daily	Weekly	Monthly	Quarterly	Annually	Never
Book	24%	60	157	241	223	148	73
Journal	87%	509	277	72	22	6	16
Preprint	18%	57	105	155	109	72	404
Conference	2%	4	14	37	193	492	162
Proceeding	5%	14	37	79	168	273	331
Web page	70%	362	277	132	67	19	45
Online database	67%	293	311	119	49	32	98
Personal communication	52%	241	228	132	114	64	123
Other	1%	5	7	3	0	2	885

Abbildung 1-4: Wie oft werden folgende Informationsquellen zur Suche verwendet?

Search tool type	Frequency	Percentage
Bibliographic/citation database	1084	47%
General Web search engine	694	30%
Fulltext digital library	156	7%
Personal search tool	125	5%
Knowledgebase Web portal	93	4%
Others	69	3%
Online or local database	52	2%
Library collection	21	1%

Abbildung 1-5: Was sind die wichtigsten Werkzeuge zur Unterstützung bei der Suche?

1.4 Aufgaben des Information Retrievals

Die eingeführten Modelle des Information Seeking beschreiben den Auslöser für den Bedarf an IR, darin unterscheidbaren Anforderungen und kontextuelle Rahmenbedingungen. Der weitere Inhalt dieser Vorlesung wird sich maßgeblich auf die Verfahren beschränken, welche das nach Bates (vergl. Abbildung 1-3) definierte *Suchen* (engl. searching) umsetzen**.** Also, den User darin unterstützen, sein aktives Bestreben Antworten zu offenen Fragen zu finden zu unterstützen bzw. ein Verständnis für einen neuen Sachverhalt zubekommen.

In der IR Literatur hat sich für die Umsetzung der Suche auch die Bezeichnung des Ad Hoc Retrievals eingebürgert. Das Ad Hoc Retrieval wird als eine der übergeoordneten *Aufgaben* (engl. Retrieval Task) im IR behandelt. Manning et al. (vergl. [MRS08]) beschreiben den Charakter einer Retrieval Task als: „*the task executed by the information system in response to a user request. It is basically of two types: ad hoc and filtering*" . In einer Retrieval Task werden also gemeinhin die beiden Aufgaben des *Ad Hoc Retrievals* und des *Filterings* unterschieden:

- Das *Ad Hoc Retrieval* ist die Standard Aufgabe des IR: Der User eines IR Prozesses spezifiziert hierbei seinen Informationsbedarf mittels einer Suchanfrage, welche ein automatisiertes Verfahren startet. Die Ausgabe des Verfahrens besteht dabei in einer klassischen IR Anwendung in einer Menge relevanter Dokumente mit textuellem Inhalt. Die Bezeichnung *Ad Hoc* leitet sich daraus ab, da die Anzahl möglicher Fragen fast unbegrenzt sind.
- Das *Filtering* beschreibt die Aufgabe in dem ein relativ statischer Informationsbedarf besteht, jedoch der Dokumentbestand sich beständig erweitert (z.B. das durchsuchen von Emaillisten).

Diese Vorlesung beschäftigt sich ausschließlich mit Themen die dem Search nach Bates, bzw. dem Ad Hoc Retrieval zugeordnet werden können. Information Filtering ist kein Bestandteil dieser Vorlesung.

Als die klassische Aufgabe des IR wird die Suche nach Volltextdokumenten verstanden. Natürlich ist das Prinzip des Ad Hoc Retrievals aber auch für die Verarbeitung multimedialer Inhalte anwendbar. Seien es Videos, Audiodateien, wissenschaftliche Analysedaten oder anderes mehr. Die Grundlagen des IR lassen sich allerdings am besten über IR Prozesse einführen die auf umfangreiche Dokumentkollektion mit textuellen Inhalten arbeiten. Historisch bedingt sind viele der grundlegenden Lösungsvorschläge des IR auf dieser Basis gewonnen worden. Im Verlauf des Skriptes bilden daher auch textuelle Dokumente einer Dokumentkollektion die Basis weiterer Erklärungen.

Neben dieser noch sehr generellen Beschreibung der Suche bzw. des Ad Hoc Retrievals in einer Dokumentkollektion, gibt es mannigfaltige spezielle Aufgabenbereiche für deren Anwendung. Aufgrund der hohen Anzahl an unterschiedlichen Aufgabenbereichen, soll an dieser Stelle nur

eine Auswahl eingeführt werden. Die Beschreibungen sind den sog. *Tracks* einer etablierten IR Evaluationskonferenzen TREC (*Text REtrieval Conference*) entnommen worden:

- *Complex Answer Retrieval Track*: Entwicklung von Systemen die es ermöglichen einen komplexen Informationsbedarf aufzulösen, indem relevante Informationen von einem vollständigen Korpus zusammengeführt werden.
- *Conversational Assistance Track*: Entwickeln und Testen von Systemen für automatisierte Dialoge.
- *Decision Track*: Entwickeln und Testen von Entscheidungsunterstützenden Systemen.
- *Incident Streams Track*: Entwickeln und Testen von Systemen die Social Media Daten für Notfallsituationen aufarbeiten.
- *News Track*: Entwickeln und Testen von Systemen die in Nachrichtenmeldungen arbeiten.
- *Precision Medicine Track*: Entwickeln und Testen von Systemen die Verbindungen zwischen klinischen Tests und evidenzbasierter Literatur herstellen, um an neueste effektive Behandlungsformen zu finden.

Eine weitere Unterteilung im IR betrifft die Art und Weise in der das Retrieval umgesetzt wird und welche Voraussetzung seitens des Users vorausgesetzt werden. Zu unterscheiden sind hierbei zwei Klassen. Die Ansätze der sog. *Exact Match* Klasse erwarten das der User ein genaues Wissen über den Inhalt und Aufbau einer Dokumentensammlung hat. In diesem Fall kann der User über eine wohldefinierte Anfragesprache seinen Bedarf an Information decken. Dazu diametral stehen die Ansätze die der sog. *Best Match* Klasse zugeordnet werden. In dieser Klasse wird davon ausgegangen, dass der User weder ein vollständiges Wissen über den Aufbau und Inhalt des Datenbestandes hat noch eine vollständig differenzierte Vorstellung darüber wie er seinen Bedarf an Informationen ausdrücken kann. In dieser Vorlesung werden maßgeblich Lösungen zu Best Match Modellen behandelt. Lösungen zu Anforderungen des Exact Matches werden eingeführt aber nicht in der Tiefe behandelt.

Der Auszug an möglichen Aufgaben die über ein IR-Prozess automatisiert werden können zeigt, dass die Anforderungen an Lösungen sehr breit angelegt sind. Es ist nicht ausreichend Dokumente nur nach einem Bestimmten Verfahren auszuwählen. Vielmehr muss es Lösungen bieten den Datenbestand für verschiedene Aufgaben vorbereitet zu können, auch die Bandbreite an möglichen Ausgabeformaten zu unterstützen und die Interaktion mit dem User umsetzen. Aufgrund der Vielzahl der Anforderungen soll im nächsten Kapitel ein sehr generelles konzeptuelles IR-Modell vorgestellt werden.

1.5 Generelles, konzeptuelles IR Modell

Allgemeingesprochen kann das Ad Hoc Retrieval als ein Prozess aufgefasst werden. Dieser Prozess beinhaltet eine bestimmte Abfolge von Verarbeitungsschritten, auf einer vorverarbeiteten textuellen Dokumentmenge. Ein generelles Modell, welches diesen Prozess in seiner Gesamtheit abbildet ist u.a. von N. Fuhr in [F92] eingeführt worden und soll stellvertretend für weitere Modelle hier dokumentiert werden (vergl. auch Abbildung 1-6).

Zunächst wird festgelegt, dass ein Dokument mit der Variablen dn bezeichnet wird, und $d_n \in D$ gilt. Eine Query wird mit q_k bezeichnet und es gilt $q_k \in Q$. Es wird weiterhin festgelegt, dass es zwischen d_n und q_k eine Relevanzbeziehung gibt. Z.B.: $\mathcal{R} = \{R, \bar{R}\}$, wobei R bezeichnet, dass eine Dokument zur Query relevant ist und $\bar{R}$, dass das Dokument zur Query nicht relevant ist. Diese Festlegung impliziert insbesondere die Gültigkeit der Abbildung $r: Q\ x\ D \rightarrow \mathcal{R}$.

Ein IR Prozess hat jedoch nur eingeschränkte Möglichkeiten ein Dokument oder eine Query darzustellen. Daher wird bei der Verarbeitung von d und q auch von einer abstrahierten Darstellung $d'_n \in D'$ und $q'_k \in Q'$ ausgegangen, die durch die Abbildungen $\alpha_D$und α_Q erzeugt wird. Eine abstrahierte Darstellung könnte z.B. eine Menge von Wörtern des Dokuments und der Query der Betrachtung sein. Ähnlich dem Index eines physischen Buches, welcher auch den Inhalt über enthaltenen Schlagwörter repräsentiert. Fuhr geht in seinem Modell noch einen Abstraktionsschritt weiter, denn es gibt IR Modelle, welche auf einer weiteren Abstraktionsstufe von Dokumenten und Queries arbeiten. Die weitere Abstraktion der Darstellung zu $d''_n \in D''$ und $q''_k \in Q''$ erfolgt über die Abbildungen β_D und β_Q. Eine sog. *Retrievalfunktion* berechnet anschließend die Relevanz zwischen Dokument- und der Query-Repräsentation: $\rho(q''_k, d''_n)$. Das Modell, welches der Retrievalfunktion zugrunde liegt, wird als das *Retrievalmodell* bezeichnet. Das Ergebnis der Berechnung hingegen ist ein Wert der die Relevanzbeziehung ausdrückt und wird gemeinhin als der *Retrieval Status Value* (RSV) bezeichnet.

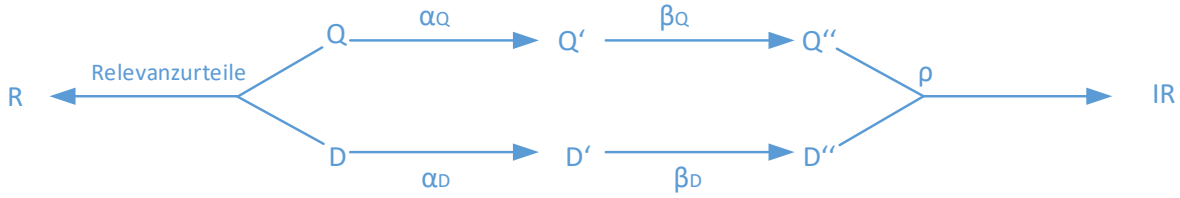

Abbildung 1-6: Konzeptuelles IR Model

1.6 Dokumente, eine Wissensrepräsentation für den Informationsaustausch

Nach Belkins ASK-Modell entsteht der Informationsbedarf eines Menschen aus einem Zustand des nicht vorhandenen Wissens. Dieses fehlende Wissen soll über einen IR-Prozesses kompensiert werden. Der IR-Prozess muss folglich auf eine „Menge an Wissen" zugreifen, eine relevante Untermenge daraus selektieren und als Resultat der Suche an den User des IR-Prozesses ausgeben. In den vorangegangenen Kapiteln wurde bereits beschrieben, dass potentiell alle digitalen multimedialen Datensätze Wissen repräsentieren können. In diesem Skript, beschränken wir uns jedoch explizit auf textuelle Inhalte um Wissen zu repräsentieren. Nach Abbildung 1-6 und den weiteren bereits eingeführten Konzepten des IR, wird dieser Kurs insofern diskutieren, wie die Repräsentation eines Informationsbedarfs, mit der Repräsentation eines textuellen Dokuments verglichen wird, um einen RSV zu berechnen auf dessen Basis eine sortierte Ausgabe einer Suche erfolgen kann. Das vorangegangene Kapitel hat bereits eingeführt, dass ein Dokument zunächst in eine passende Repräsentation überführt werden muss. Der Prozess, ein digitales Dokument in eine für das IR passende Repräsentation umzuwandeln, wird hier als das *Indexieren* bezeichnet. Damit jedoch der Indexierungsprozess, ein Dokument in eine Dokumentrepräsentation transformieren kann, soll zunächst ein Grundverständnis für den generellen Aufbau digitaler textueller Dokumente besprochen werden. Obwohl die einzelnen Bestandteile eines textuellen Dokuments eigentlich bekannt sein sollten, sollen diese hier noch einmal explizit benannt werden.

Die gängige Methode um Wissen, für den Austausch von Informationen in einem Dokument festzuhalten ist die Verwendung der natürlichen Sprache. Um die natürliche Sprache festzuhalten wird die Schriftform verwendet, welche die Sprache in Form von Text darstellt. Ein Text wiederum setzt sich aus Wörtern einer bestimmten Sprache (z.B. Deutsch, Englisch oder Französisch) zusammen der für die spätere Verarbeitung auf einem Trägermedium festgehalten werden muss. Um solche Texte auf Trägermedien für einen Computer lesbar kodieren bzw. sie für die Verarbeitung über ein IR-Prozess aufzubereiten, werden weitere Vereinbarungen benötigt. Dies sind z.B. die Einhaltung von Dateiformaten (z.B. PDF oder DOCX), welche wiederum auf einer Vielzahl an Modellen und Standards basieren. Ein Text der zu diesen Vereinbarungen Konform ist, wird als Dokument bezeichnet. Ein Vorlesungsskript wie dieses oder eine wissenschaftliche Publikation sind gute Beispiele für eine schriftliche Wissensmanifestation aus natürlicher Sprache in Form eines Dokuments.

Zu beachten ist, dass bei Textdokumenten die über ein Dateiformat spezifiziert sind, über den reinen Text hinaus auch weitere Auszeichnungen enthalten sein können, die einem Text eine Semantik zuweisen. Bei einer wissenschaftlichen Publikation kann dies z.B. die Unterteilung in Kapitel und Unterkapitel sein. Jedes Kapitel kann wiederum aus einer Menge an Absätzen bestehen, welche Listen, Bilder, Tabellen oder ähnliche Konstrukte beinhalten. Ein weiteres Beispiel wären XML Dokumente, die neben dem reinen Text noch Auszeichnungen beinhalten, welche den Text semi-strukturiert in sinnhafte Abschnitte unterteilen bzw. mittels Attribute Text um weitere Bedeutung erweitern. Oder XHTML Dokumente, welche zusätzliche Steueranweisungen

für einen Browser (z.B. </br>, <body> oder <head>) aufweisen. Weiterhin zu erwähnen ist das Text natürlich auch in klassischen Datenbank-Anwendungen, wie relationalen Datenbanken abgelegt sein kann. Solche Textobjekte sind insofern keine trivialen Objekte, sondern können hinreichend komplex sein und reinen Text durch weitere Bedeutung oder Anweisungen anreichern. Im weiteren Verlauf dieses Skriptes gehen wir davon aus, dass ein Dokument allgemein drei Eigenschaften teilt:

- ***Inhalt:*** z.B. Text, Bilder oder Tabellen.
- ***Struktur:*** z.B. Kapitel, Titel und Absätze.
- ***Layout:*** z.B. Schriftgröße, Farbe oder Einrückungen.

Nachdem der generelle Aufbau eines Dokuments beschrieben wurde, verbleibt die Frage wie der eigentliche textuelle Inhalt kodiert wird. Da es viele Mögliche Modelle zur Kodierung von Text gibt, soll an dieser Stelle das das generelle Modell kurz umrissen werden. Das Modell besteht aus fünf aufeinander aufbauenden Ebenen, die in Form der folgenden Liste eingeführt werden:

- ***Ebene 1, der abstrakte Zeichensatz***: Die (ungeordnete) Menge aller verwendbaren abstrakten Zeichen. Buchstaben, Ziffern, Interpunktionszeichen, Akzente, grafische Symbole, ideografische Zeichen, Leerzeichen, Tabulatoren, Zeilenweiterschaltung und -vorschub und Kontrollkodes für Auswahl-Markierungen. Der Zeichensatz ISO 646-US (US-ASCII) beispielsweise besteht aus 33 Kontrollzeichen und 95 druckbaren Zeichen.
- ***Ebene 2, die Codetabelle***: ist eine Zuweisung von Abstrakten Zeichen zu einer Codeposition. Eine natürliche Zahl größer null. Diese Codepositionen werden als hexadezimal Zahlen angegeben, um eine Beziehung zu Bitmustern darzustellen.
- ***Ebene 3, das Kodierungsformat***: Mithilfe des Kodierungsformats für eine Codetabelle werden die Bitrepräsentationen für die Codeposition festgelegt. Dazu gibt es eine Code-Einheit, die in der Regel aus acht oder sechzehn Bits besteht. Durch das Kodierungsformat werden dann die Positionen eines Code-Raums in Sequenzen von Code-Einheiten und somit in Bitmuster abgebildet. Wird jede Codeposition einer Codetabelle auf die gleiche Anzahl von Code-Einheiten abgebildet, sprechen wir von einem Kodierungsformat fester Länge; andernfalls hat das Kodierungsformat variable Länge. Ein Beispiel Kodierungsformat variabler Länge UTF-16.
- ***Ebene 4, das Kodierungsschema:*** Damit Daten zuverlässig über ein Netzwerk ausgetauscht werden können, müssen sie in eine lineare Folge von Bytes gebracht werden. Dieser Vorgang wird auch *Serialisierung* genannt. Mithilfe eines Kodierungsformats wird der Text als Folge von Kodierungseinheiten dargestellt. Ein Kodierungsschema wird nun dazu genutzt zu einem Kodierungsformat zusätzlich festzulegen, wie die Kodierungseinheiten in Bytefolgen zu serialisieren sind. Ist die Kodierungseinheit eines Kodierungsfor-

mats selbst acht Bits lang, so ist nichts mehr festzulegen und das Kodierungsschema ist mit dem Kodierungsformat identisch. So kann also UTF8 sowohl als Kodierungsformat als auch als Kodierungsschema bezeichnet werden. Ist die Kodierungseinheit ein Word, wie bei UTF16, so gibt es zwei Möglichkeiten der Serialisierung: Big Endian1und Little Endian. Dementsprechend gibt es zu UTF16 zwei Kodierungsschemata: UTF16-BE und UTF16-LE. Analoges gilt für UCS2 und UCS4.

- ***Ebene 5, die Syntax:*** Mit glattem Text, der als Folge von Unicode-Zeichen kodiert ist, kann der reine Inhalt von strukturierten Dokumenten gehandhabt werden. Da auch eingebettetes Markup – wie etwa bei den in Abschnitt 1.3.7 eingeführten Tags der Form <XXX> und </XXX> – eine Folge von Zeichen ist, scheint es auf den ersten Blick als wäre das Problem, strukturierte Dokumente zu kodieren, bereits vollständig gelöst. Es ergibt sich jedoch eine zusätzliche Komplikation dadurch, dass der Markup-text vom Inhaltstext syntaktisch voneinander unterscheidbar sein muss. Zur Abgrenzung von Markup und Inhalt werden bestimmte Funktionszeichen eingesetzt, die dann außerhalb ihrer syntaktischen Rolle weder im Markup selbst noch im Inhalt des Klartextes vorkommen dürfen. In XML ist „<" ein solches Funktionszeichen, das den Anfang eines Tags charakterisiert. Das Zeichen „<" darf deswegen im Inhaltstext eines XML-Dokuments nicht vorkommen. Die klassische Methode Funktionszeichen in glattem Text unterzubringen, ist die Verwendung von so genannten Escape-Zeichen. Diese werden nicht selbst als Bestandteil des Textes interpretiert sondern signalisieren die Präsenz eines Zeichens, das eigentlich nicht im Text vorkommen darf. Im Falle von XML bezeichnet die Formel &#<<Hexziffern>>; oder als &<<Dezimalziffern>>; im Inhaltstext das Unicode- Zeichen an Position <<Hexziffern>> bzw. an Position <<Dezimalziffern>>. Wir können ein „<" im Inhaltstext als z. B. als < notieren und das zusätzliche Funktionszeichen „&" durch &.

Wie eingangs des Kapitels erwähnt, ist die natürliche Sprache in Form von Text aber bei weitem nicht die einzige Form in der sich Wissen maschinenlesbar darstellen lässt. IR-Prozesse können auch auf Audiosignalen bzw. auch auf visuelle Strukturen wie Bilder, Grafiken oder ähnliches basieren. Auch sie können digitalisiert und damit für den Computer bearbeitbar aufbereitet werden. Das gros der ist jedoch immer noch mit der Verarbeitung von textbasierten Manifestationen von Wissen in Form von Dokumenten beschäftigt

Auch sei an dieser Stelle darauf hingewiesen, dass der Informationsbedarf eines Users nicht in jedem Fall mittels eines kompletten Dokuments adressiert werden kann. In vielen Fällen sind nur bestimmte Einheiten, wie Kapitel oder Absätze eines Dokumentes von Interesse für den User und andere eher gar nicht. U.a. im Kontext von Suchen die über das Internet verarbeitet wer-

den, können dabei drei unterschiedliche Ausprägungen von Bedürfnissen unterschieden werden, die unterschiedliche Bestandteile von Web-Dokumenten adressieren:

- **Informational Queries:** Queries bei denen der Informationsbedarf durch ein breitgefächertes Ergebnis adressiert wird. Bei dieser Art von Bedürfnis werden generelle Informationen gesucht.
- **Navigational Queries:** das Bedürfnis wird am besten durch eine ganz bestimmte Website bedient. Eine breitangelegte Ergebnismenge ist nicht gewünscht. Dies können z.B. die Website eines Konzerthauses oder einer Klinik sein.
- **Transactional Queries:** das Bedürfnis des Users bezieht sich auf durch Durchführung einer Kauftransaktion. Die Ergebnismenge besteht im besten Fall aus Formularen, über welche eine Transaktion umgesetzt werden kann.

1.7 IR Systeme

Das Ziel eines IR Prozesses ist die Suche nach Informationen automatisiert zu unterstützen. Bislang haben wir dabei immer von einem IR Prozess gesprochen. Systeme die einen IR Prozess, wie in Abbildung 1-6 gezeigt automatisieren, werden *Information Retrieval System* (IRS) genannt und sind eine spezielle Ausprägung eines IS. Auch zu dem Thema IS gibt es viele unterschiedliche Ansichten und Definitionen derer Funktionsweise. Dieses Kapitel gibt einen Überblick über bestehende Definition und Modelle, ausgehend von Modellen aus dem Umfeld der IS wird dann ein konzeptuelles Modell eines IRS eingeführt.

Zunächst soll aus den Definitionen des Kapitels 1.2.1 folgender Zusammenhang rekapituliert werden: U.a. in den Wirtschaftswissenschaften sind Daten die Basis um Information abzuleiten. Aus abgeleiteter Information manifestiert sich Wissen. In den Wirtschaftswissenschaften ist dabei ein häufig referenziertes Modell welches diese Zusammenhänge beschreibt, die sogenannte *Daten Information Wissen und Erfahrungs-Hierarchie* (engl. Data–Information–Knowledge–Wisdom Hierarchy (DIKW-Hierarchy)) siehe z.B. [R07]. Ersichtlich aus diesem Modell ist insbesondere, dass wie bereits eingeführt *Information* auf dem Zugriff auf *Daten* beruht. Als Daten werden dabei Werte aufgefasst, die unverarbeitet und diskret sind (vergleiche Abbildung 1-7).

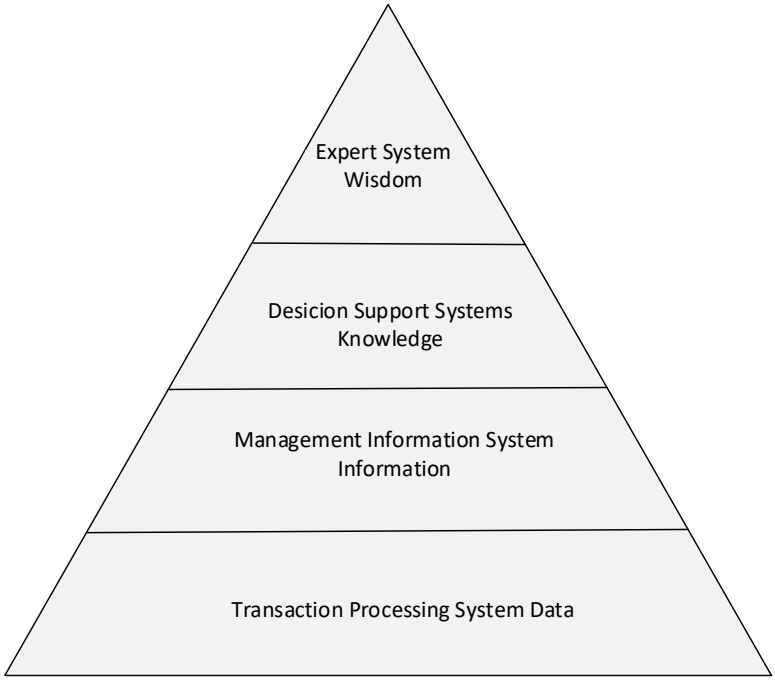

Abbildung 1-7: Data - Information - Knowledge - Wisdom Hierarchy nach [R07]

Die Struktur der über ein IS zugreifbaren Daten wird über ein *Datenmodell* beschrieben. Kemper et al. definieren dies in [KE11] als: *„Das Datenmodell legt die Modellierungskonstrukte fest, mittels der man ein computerisiertes Informationsabbild der realen (bzw. des relevanten Ausschnitts) der Welt generieren kann" „[…] es legt die generischen Strukturen und Operatoren fest die man zur Modellierung einer bestimmten Anwendung ausnutzen kann.".*

Ein IS lässt sich damit insofern als ein System beschreiben, dass auf Basis von Daten, die auf der Spezifikation eines Datenmodells beruhen, Information für den Zugriff eines Users bereitstellt (vergleiche Abbildung 1-8) oder analog zu dieser Beschreibung: *„An information system is a formalized computer information system that can collect, store, process, and report data from various sources to provide the information necessary for managerial decision making"* [H93].

Abbildung 1-8: Einfaches Informationssystem-Modell nach [ACM19]

Weitere Definitionen über Informationssysteme beinhalten insbesondere zusätzlich die Aussage, dass ein solches System nicht isoliert ist, sondern Verbindungen zu weiteren Elementen der Umgebung hat, in der es eingesetzt wird. Unter anderem beschreiben Huber et al. in [SHPL04] ein Informationssystem als eine Sammlung aus *Menschen, Information, Geschäftsprozessen* und *Informationstechnologien*, die darauf ausgelegt ist, Eingaben in eine bestimmte Ausgabe zu transformieren, um damit bestimmte Geschäftsziele zu erreichen. Auch Hevner et al. sehen ein Informationssystem nicht in Isolation. Sie definieren in [HMPR04] ein Informationssystem als eine Komposition aus *Menschen, Strukturen* und *Technologien*. Johnston et al. schließen in [JBT04], dass jeder Prozess aus Abbildung 1-8 als eine Menge von verknüpften Subsystemen abgebildet werden kann, die je nach Detaillierungsgrad weiter aufgeschlüsselt werden (vergleiche Abbildung 1-9).

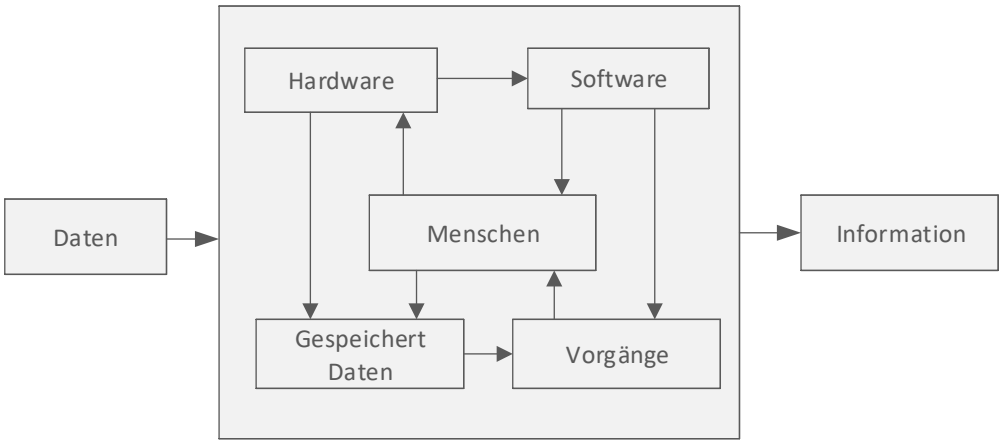

Abbildung 1-9: IS nach Schultheis in [JBT04]

Ein IRS ist eine spezielle Ausprägung eines IS, zur automatisierten Suche in Dokumentkollektion gegeben eines Informationsbedürfnisses. Die bislang eingeführten IS Modell nehmen jedoch eine Wirtschaftswissenschaftliche orientierte Sicht auf ein IS ein. Wie bereits im Kapitel 1.2.1 beschrieben etabliert dieser Kurs die Sichtweise von Kuhlen auf die Zusammenhänge zwischen Daten Wissen und Informationen. Kuhlen, dessen semiotische Sicht im Gegensatz zur wirtschaftswissenschaftlichen Sichtweise einen Zusammenhang von Daten als Grundlage von Wissen, als Grundlage von Informationen vorsieht, fasst die automatisierte Transformation von Wissen in Information im subjektivem Kontext des Users auf. Hierbei sind kontextuelle

Rahmenbedingungen und der aktuelle Wissensstand eines Users Antrieb für die Verwendung eines IRS umsetzt.

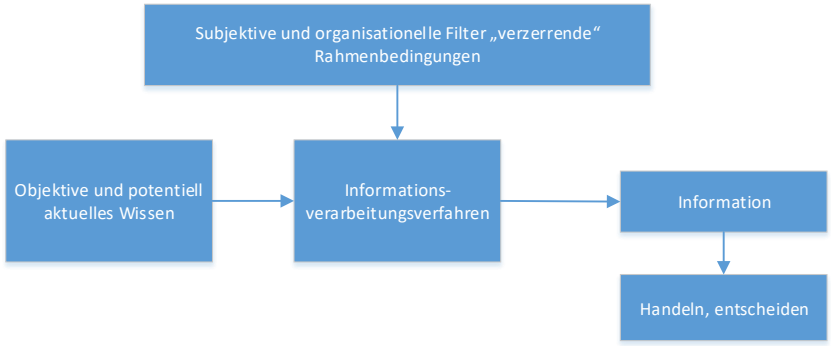

Abbildung 1-10: Transformation von Wissen in Information [K85]

Die folgend eingeführte Sicht auf ein konzeptuelles IRS ist der Publikation von Buckland et al. (vergl. [BP94]) entnommen. Das konzeptuelle Modell sieht generell eine Unterteilung wie die in Abbildung 1-8 vor. Es gibt eine Eingabe in das System, das System hält eine Menge an Verarbeitungskomponenten vor, die eine Ausgabe erzeugen. In Abbildung 1-8 ist das Vorgehen prinzipiell dargestellt. Kästchen der Abbildung die keinen gestrichelten Rahmen haben beschreiben allg. Systemfunktionalitäten. Die Bedeutung der einzelnen Komponenten wird in den folgenden Kapiteln, anhand der benannten Reihenfolge erläutert.

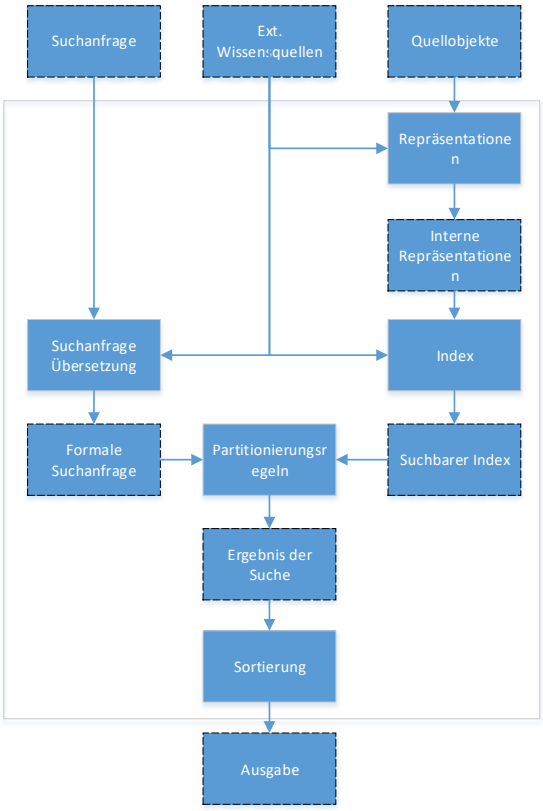

Abbildung 1-11: konzeptuelles IRS Modell nach [BP94]

1.7.1 Systemeingabe

Die Eingabe in das IRS kann in drei Bereiche unterteilt werden. Der erste Bereich ist der, welcher die Query eines Users beschreibt. Eine gängige Form der Übermittlung einer Query wird über die Angabe von Zeichenketten vorgenommen. Wobei die Zeichenketten Bestandteil einer Formalen Sprache (u.a. SQL oder SPARQL) sein können oder umgangssprachlich. Dies könnten aber z.B. auch im Falle des Musik Retrievals, Melodien sein die vorgesummt werden. Oder ein Bild, im Falle der Suche nach ähnlichen Bildern.

Passend zur Query, muss das IRS auf eine Dokumentsammlung zugreifen können gegen die es die Query sinnhaft abbilden kann. Diese Vorlesung ist auf die Verarbeitung von Dokumenten mit Volltexten beschränkt. Potentiell könnten die Elemente der Sammlung aber jeglichen multimedialen Inhalt haben.

Neben der Query und den zu durchsuchenden Elementen wird zusätzlich die Berücksichtigung externer Wissensquelle benannt. Diese externeren Wissensquellen können bestimmte Datenbanken, aber auch menschliche Expertise sein. Dass über diese Quellen verfügbare Wissen wird u.a. genutzt um Queries zu optimieren (z.B. in dem die Bedeutung eines Wortes einer umgangssprachlichen Query festgelegt wird. Z.B. „Jaguar", Automarke oder Tier?). Auch wird dieses Wissen genutzt um den glatten Text eines Dokuments der Dokumentsammlung in eine für das Retrieval verwertbare Form zu überführen. Dieser Vorgang wird als Indexieren bezeichnet.

1.7.2 Verarbeitungskomponenten

Die Verarbeitungskomponente bekommt die Eingabe und verarbeitet diese für unterschiedliche Zwecke. Bevor eine Query überhaupt verarbeitet werden kann, müssen jedoch erst alle Elemente der Dokumentsammlung in eine für die Retrievalfunktion verständliche Repräsentation überführt worden sein. Der sog. Indexierungsvorgang untersucht dafür den Inhalt der Elemente der betrachteten Dokumentsammlung und überführt diesen, unter Umständen unter Zuhilfenahme externen Wissens in eine entsprechende Repräsentation (den Index). Der Index wird in einen für das IRS zugängliches Speichermedium abgelegt.

Auch die Query muss gegebenenfalls um externes Wissen angereichert werden und dann in eine für die Retrievalfunktion verständliche Repräsentation überführt werden. Die Retrievalfunktion bekommt Repräsentationen der Query und der Elemente von Dokumenten und vergleicht diese hinsichtlich Ihrer Relevanz. Gegebenenfalls wird das berechnete Ergebnis noch sortiert und danach an die Systemausgabe weitergegeben.

1.7.3 Systemausgabe

Die Systemausgabe visualisiert das Ergebnis der Verarbeitungskomponente für den User und reichert dieses gegebenenfalls noch um weitere Informationen an.

1.8 Ausblick

Die vorangegangenen Kapitel haben Grundlagen des IR diskutiert, die im weiteren Verlauf des Kurses vertieft werden sollen. Die Abbildung 1-12 ist ein weiteres Schaubild, dass eine abstrakte Sicht auf ein IRS darstellt, welche sich aus den vorangegangenen Modellen ableiten lässt. Zusätzlich zu den Bestandteilen eines IRS stellt es eine Verbindung zu den folgenden Kapiteln dieses Skriptes her und wird als Orientierungshilfe in den kommenden Kapiteln KE2 – KE7 eingeführt werden.

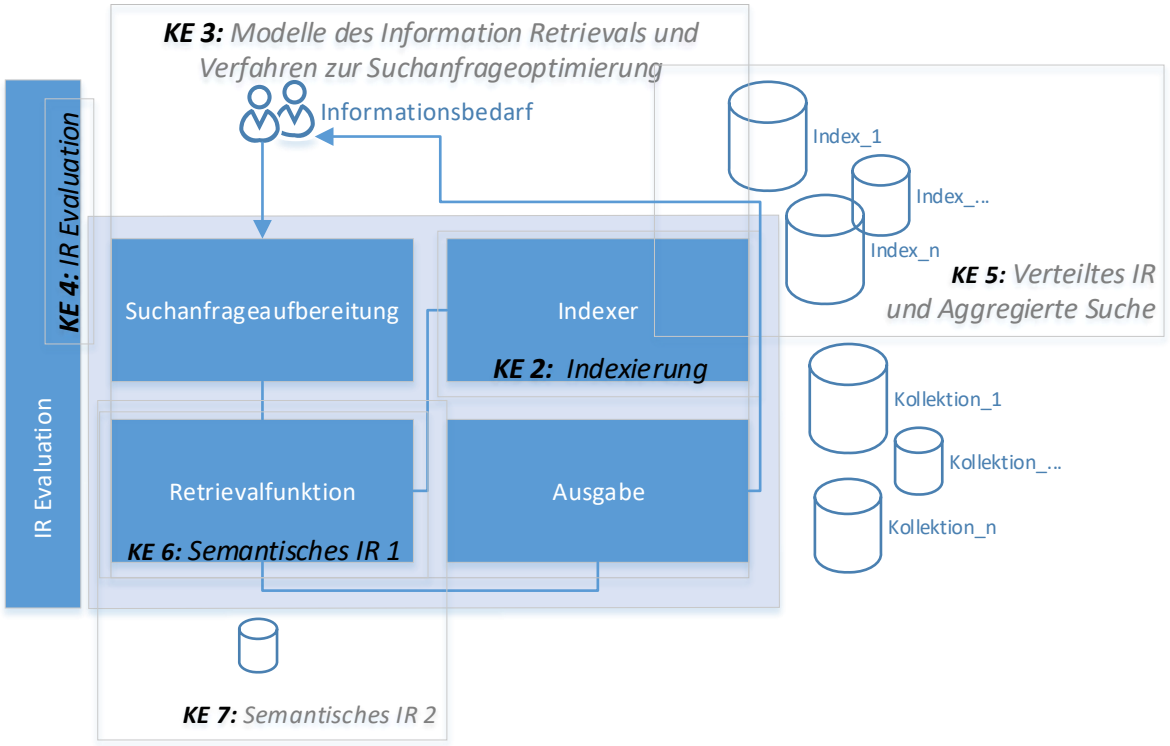

Abbildung 1-12: Zuordnung Kursaufbau / Funktionale IR Eigenschaften

1.9 Selbsttestaufgaben

Achtung, die Selbstestaufgaben werden nicht von uns korrigiert. Sie dienen Ihrer selbstständigen Vertiefung der Inhalte.

1.9.1 Schriftliche Selbstestaufgaben

- **Kapitel 1.2**
  - Geben Sie in Ihren eigenen Worten wieder was Kuhlen als *Information* definiert.
  - Welcher Zusammenhang besteht zwischen Daten, Informationen und Wissen nach Kuhlen?
- **Kapitel 1.3**
  - Was ist ein Informationsbedarf und welche Rolle spielt er im IR?
  - Was ist der Unterschied zwischen Information Retrieval und Information Seeking?
- **Kapitel 1.4**
  - Welche Information Retrieval Aufgaben wurden vorgestellt?
  - Welche Information Retrieval Anwendungen haben Sie kennengelernt? Welche weiteren gibt es?
- **Kapitel 1.5**
  - Was ist eine Dokument- und eine Query-Repräsentation?
  - Was ist der RSV?
- **Kapitel 1.6**
  - Welche generellen Eigenschaften eines Dokumentes wurden vorgestellt?
  - Welche Query Arten haben Sie kennengelernt? Wie unterscheiden sich diese?
- **Kapitel 1.7**
  - Beschreiben Sie in ihren eigenen Worten was eine Retrieval Funktion ist. Bitte berücksichtigen Sie darin zu beschreiben, welche Verbindung Sie zum Indexierung sehen.
  - Welche Rolle spielt der Index beim Information Retrieval?

1.9.2 Java Programmieraufgaben

Achtung, die Java Programieraufgaben werden, wie die schriftliche zu verfassenden Selbstestaufgaben nicht von uns korrigiert! In den folgenden Kurseinheiten können Sie auch ganz praktische Erfahrungen mit bestehenden IR Softwarebibliotheken erlangen. Die Anwendung des erlernten dient der Vertiefung Ihres Verständnisses der Materie und dem praktischen Nutzen der vorgestellter Verfahren. Dazu werden wir Ihnen einige Java Programmie-

rübungen zur Verfügung stellen. Da diese praktische Übung kein Bestandteil bei der Anmeldung zum Kurs war, und die Studenten unterschiedliche Kenntnisse in der Programmierung haben. Soll hier zunächst etwas ausführlicher ausgeholt werden.

Um einen einheitlichen Einstieg in diesen praktischen Übungsbetrieb zu erlauben, wird zunächst eine einheitliche Umgebung auf Ihrem Computer eingerichtet. Die Einrichtung umfasst die folgenden Schritte: Installation der *Java JDK*, einer Entwicklungsumgebung (*IntelliJ*) und der Einrichtung des Übungsprojektes (*Maven*). Getestet wird die Installation über die Ausführung eines ersten kleinen Programmes zur Indexierung von Volltext.

1.9.2.1 Installation

Java JDK: Zunächst muss eine aktuelle Java ***JDK*** Installation (falls noch nicht vorhanden) vorgenommen werden. Die aktuellen JDKs von Oracle finden Sie hier[1] zum Herunterladen. Installieren Sie das JDK (eine JRE ist nicht ausreichend!). Bitte stellen Sie sicher, dass die Installation erfolgreich verlaufen ist und auch die Umgebungsvariable „JAVA_HOME" korrekt gesetzt wurde (vergl. auch hier. https://www.baeldung.com/java-home-on-windows-7-8-10-mac-os-x-linux). Dazu öffnen Sie ein Terminal und geben den folgenden Befehle ein: *java –version* Bei erfolgreicher Installation sollte sich eine folgende Ausgabe erfolgen:

```
C:\Users\Engel>java -version
java version "13.0.2" 2020-01-14
Java(TM) SE Runtime Environment (build 13.0.2+8)
Java HotSpot(TM) 64-Bit Server VM (build 13.0.2+8, mixed mode, sharing)
```

Nachdem das Java JDK erfolgreich installiert wurde benötigen Sie noch eine Entwicklungsumgebung (IDE). Wir empfehlen die freiverfügbare *Community IntelliJ Edition* (https://www.jetbrains.com/idea/download). Bitte laden Sie diese einfach herunter und installieren Sie es mit der beigefügten Installationsanleitung.

Maven[2]: Ist ein Software Managementool, dass wir benutzen um einheitliche Konfiguration des Projektes zu gewährleisten, sowie um weitere Abhängigkeiten automatisiert aufzulösen. Maven ist schon in der IntelliJ Installation integriert und bedarf keiner zusätzlichen Installation.

Besorgen Sie sich nun in Moodle, das *Kurseinheit 1: Zusatzmaterial* Basisprojekt (*Information Retrieval Kurs 1879.zip*). Dieses muss zunächst in IntelliJ importiert werden. Dazu entpacken Sie das zip Projekt zunächst und speichern es an eine Stelle Ihrer Wahl auf dem Computer. Danach öffnen Sie IntelliJ und importieren Sie das Projekt indem Sie auf „*File*" und dann auf „*open*" gehen und das Projekt auswählen (das „*Information Retrieval Kurs 1879*"- Verzeichnis in dem sich u.a. auch der *src* Ordner befindet).

---

[1] https://www.oracle.com/java/technologies/javase-downloads.html

[2] https://maven.apache.org/

Testen Sie ob die *IntelliJ* Installation funktioniert hat und das Projekt ausgeführt werden kann. Dazu öffnen Sie die Klasse „*Introduction*“, dann führen Sie ein Maven *clean* aus (Rechtsklick auf „*clean:clean*“, danach „*Run Maven Build*“, vergleiche „***A***“ Abbildung 1-13).). Im Anschluss können Sie das den Code ausführen lassen in dem Sie den grünen Pfeil aktivieren (vergleiche „***B***“ Abbildung 1-13).

Ist alles korrekt verlaufen, dann sehen Sie die in Abbildung 1-14 gezeigte Ausgabe. Falls nicht schauen Sie sich zunächst einmal das Kapitel *Trouble Shooting* (vergl. Kapitel 1.9.2.3) an.

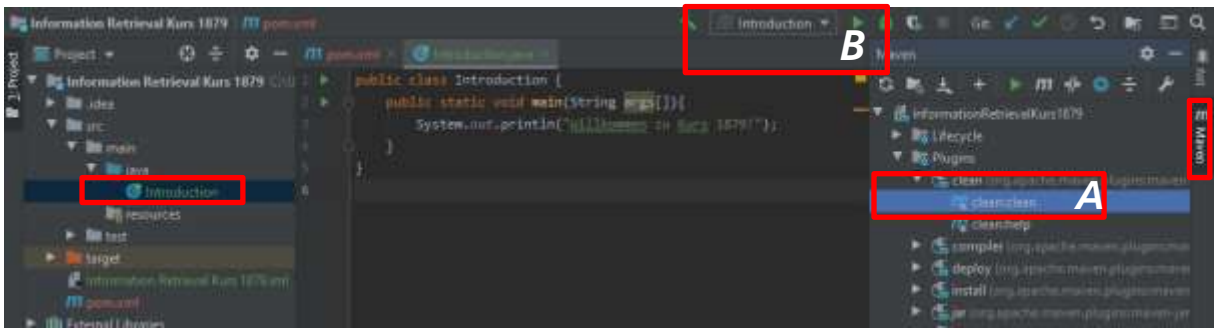

Abbildung 1-13: Maven clean

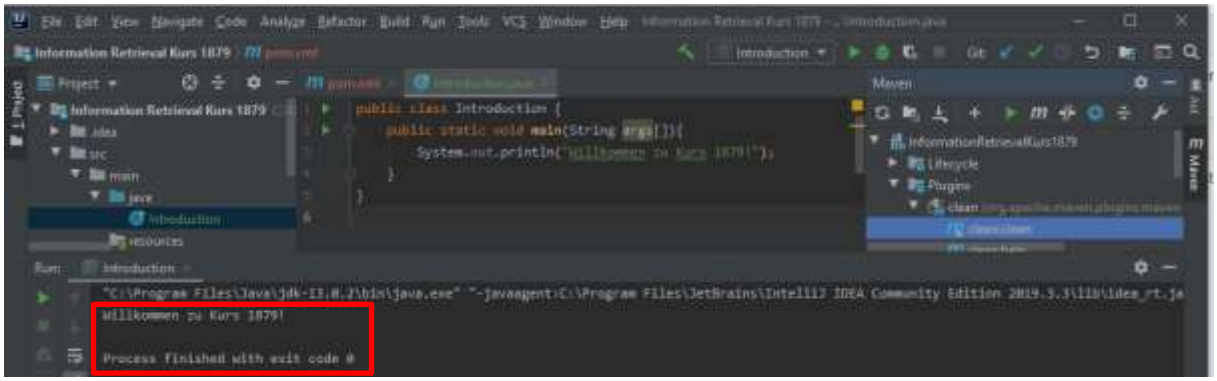

Abbildung 1-14: Ausgabe

1.9.2.2 Indexieren und Suchen in txt-Dateien

Ist die Installation von Java, *IntelliJ* und des Projektes erfolgreich vorgenommen worden. Sollen Sie nun auch schon mal ein erstes kleines Programm testen, dass eine Suchmaschine auf Basis der Lucene[3] Bibliothek implementiert und dazu einen kleinen textuellen Datenbestand indexiert. Den zu indexierenden Bestand finden Sie in dem im vorangegangenen Kapitel installierten Projekt im Verzeichnis „*main/resources/SimpleTestCollection*“ und der Index soll in dem Verzeichnis „*main/resources/index*“ erstellt werden.

[3] https://lucene.apache.org/

Um die Indexierung vorzunehmen öffnen Sie bitte in *IntelliJ* die Klasse *„de.fuh.Index"* und führen diese aus (zum Ausführen einer Klasse siehe Abbildung 1-13). Nachdem der Code erfolgreich ausgeführt wurde sehen Sie, dass dem „*„main/resources/index"* Verzeichnis einige Dateien hinzugefügt wurden. Dies ist der Index der *txt* Dateien, die Sie im Verzeichnis „*resource/SimpleTestCollection*„finden.

Lucene stellt auch eine kleine Anwendung mit dem Namen *Luke* zur Verfügung, mittels der Sie auch den soeben erstellen Index analysieren bzw. durchsuchen können. Laden Sie sich dazu die aktuelle Lucene Binary Version herunter (https://lucene.apache.org/core/downloads.html). Entpacken Sie das Projekt. Unter dem Pfad „*lucene-8.4.1\lucene-8.4.1\luke*" finden Sie dann die Luke Anwendung. Bitte starten Sie diese und geben Sie bei Nachfrage den Pfad an, unter dem Luke Ihren erzeugten Index finden kann.

Um eine Suche in Luke auszuführen machen Sie folgendes: Reiter „*Search*" hier geben Sie unter „*Query expression*" z.B. folgenden Wert ein „*vide**", dann auf den Button „Parser", dann auf den Button „Search". Achten Sie darauf, dass als „Default field" der Wert „contents" angegeben ist. Ihnen sollten nun zwei Suchergebnisse angezeigt werden.

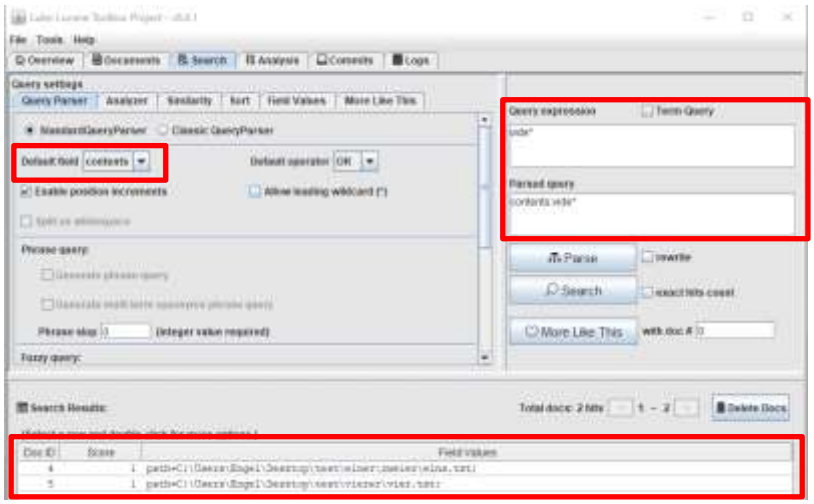

Abbildung 1-15: Ausgabe einer Suche über Lucke

1.9.2.3 Trouble Shooting

Es kann natürlich immer etwas auch nicht direkt funktionieren. Hier ein paar Hinweise die Ihnen helfen können.

Es wird das falsche JDK verwendet: Falls Sie mehr als eine JDK auf Ihrem Computer haben sollten, dann prüfen Sie ob das korrekte Java JDK verwendet wird. Dazu unter Files bitte den Eintrag „*Project Structure*" aktivieren. Es öffnet sich die in Abbildung 1-16 gezeigte Sicht. Dort fügen Sie bitte über „+" Ihr JDK hinzu.

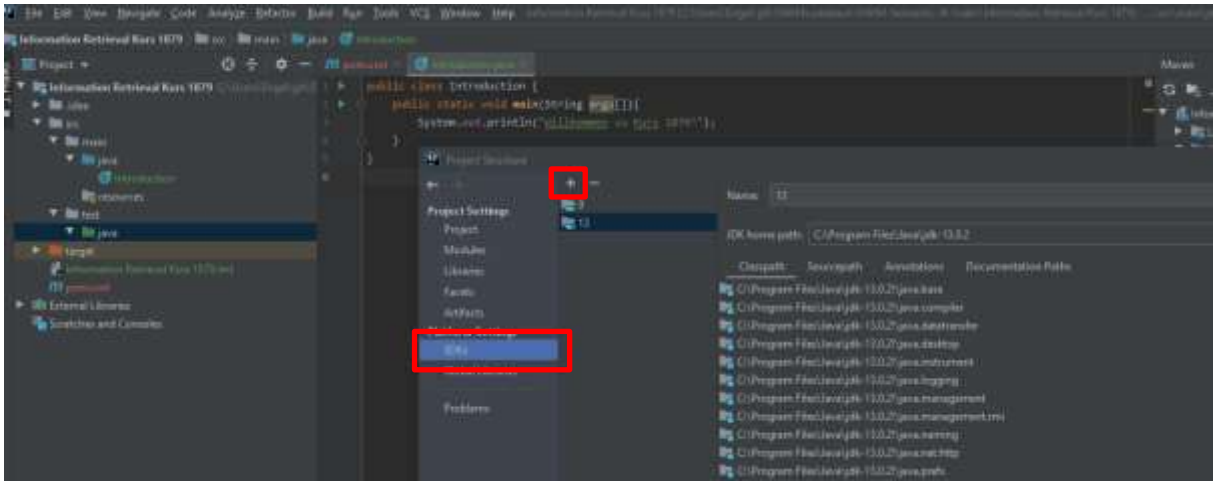

Abbildung 1-16: Project Structure

Anschließend müssen Sie noch unter Project die entsprechende IDE auswählen (vergl. Abbildung 1-17).

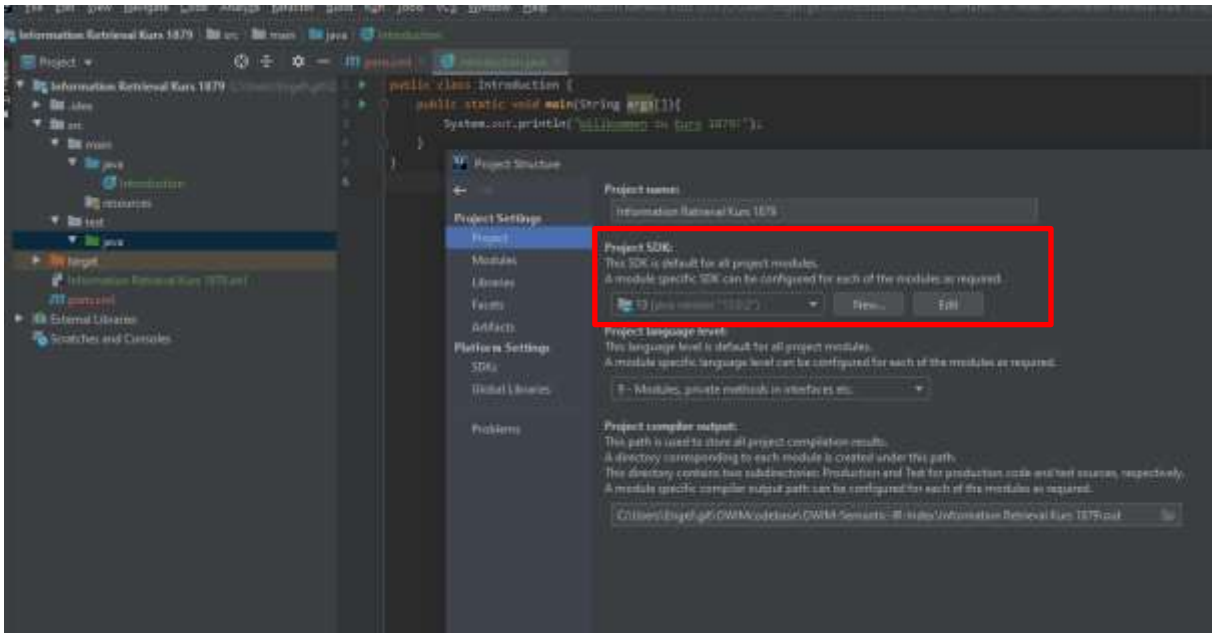

Abbildung 1-17: Auswahl der JDK

Literaturverzeichnis

[ACM19] ACM Digital Library *The ACM Computing Classification System.* https://dl.acm.org/ccs/ccs.cfm?id=0&lid=0 [2019.12.10]

[B80] Belkin, N. J. (1980). Anomalous states of knowledge as a basis for information retrieval. Canadian journal of information science, 5(1), 133-143.

[B02] Bates, M. J. (2002). Toward an integrated model of information seeking and searching. The New Review of Information Behaviour Research, 3(1), 1-15.

[BP94] Buckland, M., & Plaunt, C. (1994). On the construction of selection systems. Library Hi Tech, 12(4), 15-28.

[BYRN99] Baeza-Yates, R., & Ribeiro-Neto, B. (1999). Modern information retrieval (Vol. 463). New York: ACM press. Glossary URL: http://people.ischool.berkeley.edu/~hearst/irbook/glossary.html [2019.12.10]

[E89] Ellis, D. (1989). A behavioural model for information retrieval system design. Journal of information science, 15(4-5), 237-247.

[Dd] Duden https://www.duden.de/rechtschreibung/Daten [2019.12.10]

[F92] Fuhr, N. (1992). Probabilistic models in information retrieval. The computer journal, 35(3), 243-255.

[H93] Hicks J. 1993. Management information systems: a user perspective. Minneapolis/St. Paul : West Pub. Co., c1993, ISBN 0314933670

[HK90] Herget, J., & Kuhlen, R. (1990). Pragmatische Aspekte beim Entwurf und Betrieb von Informationssystemen. In Proceedings des 1. Internationalen Symposiums für Informationswissenschaft. Konstanz: Universitätsverlag.

[HLVA07] Hemminger, B. M., Lu, D., Vaughan, K. T. L., & Adams, S. J. (2007). Information seeking behavior of academic scientists. Journal of the American society for information science and technology, 58(14), 2205-2225.

[HMPR04] Alan R. Hevner, Salvatore T. March, Jinsoo Park and Sudha Ram (2004) Design Science in Information Systems Research Source: MIS Quarterly, Vol. 28, No. 1 (Mar., 2004), pp. 75-105 Published by: Management Information Systems Research Center, University of Minnesota

[JBT04] Johnstone, D., Bonner, M., & Tate, M. (2004) "Bringing human information behaviour into information systems research: an application of systems modelling"

Information Research, 9(4) paper 191. http://InformationR.net/ir/9-4/paper191.html [2019.12.10]

[KE11] Kemper, A., & Eickler, A. (2011). Datenbanksysteme: Eine Einführung. Oldenbourg Verlag

[K85] Kuhlen, R. (1985). Verarbeitung von Daten, Repräsentation von Wissen, Erarbeitung von Information. Primat der Pragmatik bei informationeller Sprachverarbeitung. In Sprachverarbeitung in Information und Dokumentation (pp. 1-22). Springer, Berlin, Heidelberg.

[N19] Nauta, D. (2019). The meaning of information (Vol. 20). Walter de Gruyter GmbH & Co KG.

[R07] Rowley, J. (2007). The wisdom hierarchy: representations of the DIKW hierarchy. Journal of Information Science, 33(2), 163-180.

[SHPL04] Salisbury, Wm. David; Huber, Mark W.; Piercy, Craig; and Elder, Kevin Lee (2004) "The AMCIS 2003 Panels on IS Education-I: Let Us Not Throw Out the Baby with the Bath Water: Information, Technology, and Systems All Matter in the Core IS Course," Communications of the Association for Information Systems: Vol. 14, Article 6.

[MRS08] Schütze, H., Manning, C. D., & Raghavan, P. (2008, June). Introduction to information retrieval. In Proceedings of the international communication of association for computing machinery conference (p. 260).

[UG19] University Glagow Interactive Information Retrieval Group https://www.strath.ac.uk/research/subjects/computerinformationscience/strathclydeischoolresearchgroup/ourresearchareas/interactiveinformationretrieval/ [2019.12.14]