Günter Bachelier

Polyrepräsentation, Relevanz-Approximation und aktives Lernen im Vektorraummodell des Information-Retrievals

Doktorarbeit / Dissertation

Diplom.de

Bibliografische Information der Deutschen Nationalbibliothek:

Bibliografische Information der Deutschen Nationalbibliothek: Die Deutsche Bibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://dnb.d-nb.de/ abrufbar.

ISBN: 9783832465636

Günter Bachelier

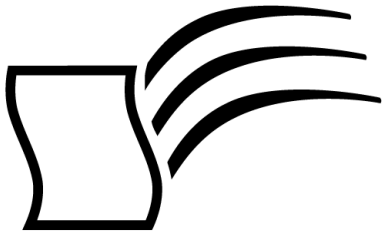
Polyrepräsentation, Relevanz-Approximation und aktives Lernen im Vektorraummodell des Information-Retrievals

Diplom.de

Günter Bachelier

Polyrepräsentation, Relevanz-Approximation und aktives Lernen im Vektorraummodell des Information-Retrievals

Dissertation / Doktorarbeit
an der Universität des Saarlandes
Fachbereich Informationswissenschaft
Juni 2001 Abgabe



Diplomica GmbH
Hermannstal 119k
22119 Hamburg

Fon: 040 / 655 99 20
Fax: 040 / 655 99 222

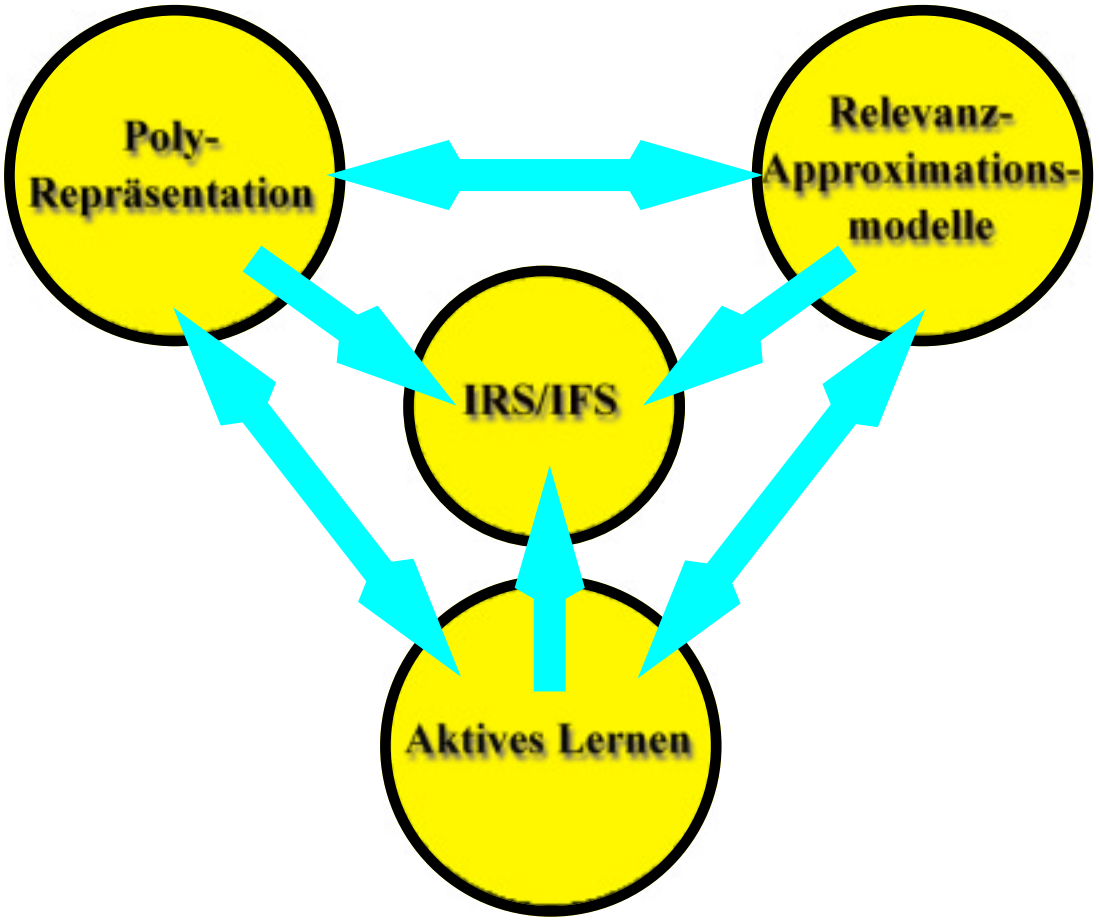
agentur@diplom.de
www.diplom.de

ID 6563

ID 6563
Bachelier, Günter: Polyrepräsentation, Relevanz-Approximation und aktives Lernen im Vektorraummodell des Information-Retrievals
Hamburg: Diplomica GmbH, 2003
Zugl.: Saarbrücken, Universität, Dissertation / Doktorarbeit, 2001

Diplomica GmbH
http://www.diplom.de, Hamburg 2003
Printed in Germany

Polyrepräsentation, Relevanz-Approximation und aktives Lernen im Vektorraummodell des Information-Retrievals



von

Günter Bachelier

Dissertation

zur Erlangung des Grades eines Doktors der Philosophie der Philosophischen Fakultäten der Universität des Saarlandes

Juni 2001

Tag der Promotion: 31.01.2002
Dekan der Philosophischen Fakultät III: Herr Univ.-Prof. Dr. Ernst Löffler
Erstberichterstatter: Herr Univ.-Prof. Dr. Harald H. Zimmermann
Zweitberichterstatter: Herr Univ.-Prof. Dr. Werner Tack

Vorwort

Konzipiert wurde diese Arbeit als elektronische Publikation in Form einer pdf-Datei mit einer Vielzahl interner und externer Links, sodass einige kurze Anwendungshinweise vorangestellt werden sollen.

Blaue, nicht unterstrichene Textbereiche repräsentieren Hyperlinks innerhalb der Arbeit, während blau unterstrichene Textbereiche auf externe Quellen (URLs) verweisen.

Eine größere Anzahl von pdf-Dateien, die im Literaturverzeichnis angegeben sind, können über die ACM Digital Library geladen werden. Der Zugang zur ACM Digital Library erfordert jedoch die Mitgliedschaft bei ACM, sowie eine Online-Anmeldung unter http://www.acm.org/dl/.

Einige Dateien, die im Literaturverzeichnis angegeben sind, besitzen den Typ XXX.djvu. Um entsprechende Dateien mit einem Browser zu betrachten, wird das Browser-Plugin “DjVuBrowser” benötigt, welches unter http://www.djvuzone.org/ geladen werden kann (Stand Mai 2001).

Die URL-Quellen beziehen sich bis auf die Angaben des Typs XXX.djvu auf den Stand April/Mai 2001.

Danken möchte ich Prof. Dr. Harald H. Zimmermann, dass ich die dargestellten Themenbereiche frei auswählen und kreativ bearbeiten konnte, sowie David Cohn für die Diskussion bezüglich einiger meiner Ideen zum aktiven Lernen (Cohn (2000[75])).

Günter Bachelier M. A.

Inhalt

1) Einleitung und Überblick

1.1) Information-Retrieval-Systeme als Spezialfall von Informationssystemen

Informationssysteme (IS) können durch einen 5- bzw. 7-Tupel bechrieben werden (Panyr (1986a: 22[247])):

$$IS = (A, W, Q, I, E). \text{ bzw. } IS = (A, W, Q, I, E, U, D). \quad (1)$$

- A: Inputfunktion (Erschließungsfunktion, Lernfunktion) zum Aufbau der internen Repräsentationen.
- W: Interne Repräsentationen (Wissensbasis).
- Q: Inputmenge als Menge aller zugelassenen Inputkonfigurationen (Problemformulierung, Suchfrage)
- I: Outputfunktion (Inferenzfunktion, Retrievalfunktion und Relevanz-Feedbackfunktion).
- E: Outputmenge als Menge aller möglichen Outputkonfigurationen (Problemlösung, Systemvorschlag).
- U: Updatefunktion der internen Repräsentationen.
- D: Dialogkomponente.

Information-Retrieval-Systeme (IRS) können als Spezialfall eines IS beschrieben werden, indem die einzelnen Komponenten des Tupels (A, W, Q, I, E) spezifiziert werden, was im Kontext des Standard-Retrieval-Prozesses in einem Vektorraummodell geschehen soll. Gegeben ist zu einem Zeitpunkt t eine Dokumentmenge D^t, die durch eine Dokument-Indexierungsfunktion $A_{IR(D)}$ auf eine Dokumentvektorenmenge DVM^t abgebildet wurde. Die einzelnen Dokumentvektoren x_i sind Element eines metrischen, n^t-dimensionalen Dokumentvektorraumes DVR, mit n^t als der Anzahl der Merkmale (Features), auf der die Indexierung basiert und die in der Menge F^t zusammengefasst werden. Der Dokumentvektorraum wird allgemein als Teilraum von $R^{n(t)}$ beschrieben, z.B. durch $[0, 1]^{n(t)}$. Die Query-Indexierungsfunktion $A_{IR(Q)}$ wird vereinfachend definiert als Abbildung aus der Menge $Q(\Theta)$ der möglichen bzw. zugelassenen Queries über einem endlichen Alphabet Θ, in DVR. Es folgt die Anwendung der Retrieval-Funktion, die abhängig ist von der momentanen Dokumentvektorenmenge DVM^t, dem Queryvektor q_i^t und dem metrischen Dokumentvektorraum DVR mit seinen definierenden Eigenschaften, wobei hier ausschließlich die Metrik d_{DVR} betrachtet wird. Sei Γ_{DVR} die Menge aller Metriken, die in einem Dokumentvektorraum DVR angewendet werden können, ohne dass hier auf die Definition der Metrik eingegangen werden soll (siehe Abschnitt 3.6.3)). Die Retrieval-Funktion kann somit spezifiziert werden als eine Abbildung der Potenzmenge $P^{DVM(t)}$ der Dokumentvektorenmenge DVM^t, dem DVR und Γ_{DVR} auf $P^{DVM(t)}$, indem das Tripel aus DVM^t, dem Queryvektor q_i^t und eine Metrik d_{DVR} auf die query-abhängige Ergebnis-Dokumentvektorenmenge DVM_i^t abgebildet wird. D.h. die Retrieval-Funktion besitzt die allgemeine Form $ret(DVM^t, q_i^t, d_{DVR})$ bzw. $ret(DVM^t, q_i^t, d_{DVR}, \varepsilon)$, wenn eine einfache Best-Match-Retrievalstrategie betrachtet wird, bei der alle Dokumentvektoren aus DVM^t selektiert werden, deren Abstand von q_i^t kleiner-gleich einer Distanzschwelle $\varepsilon \in R^+$ ist. Der letzte Schritt besteht in der Erzeugung der Dokumentmenge D_i^t, die zu der Ergebnismenge DVM_i^t korrespondiert. Vereinfachend wurde auf die Beschreibung einer Ranking-Funktion verzichtet, die aus DVM_i^t eine geordnete Liste von Dokumentvektoren erzeugt. Wird diese Darstellung des Retrievalprozesses als Grundlage verwendet, so ergibt sich die grundlegende Beschreibung eines IRS ohne eine Relevanz-Feedbackfunktion als Element von I durch das folgende Tupel:

$$IRS = ((A_{IR(D)}, A_{IR(Q)}), (D^t, DVM^t, F^t, N_{DV}{}^t), Q(\Theta), ret(DVM^t, q_i^t, d_{DVR}), \{P^{DVM(t)}, P^{D(t)}\}). \quad (2)$$

- $A = (A_{IR(D)}, A_{IR(Q)})$:	Dokument-Indexierungsfunktion, Query-Indexierungsfunktion.
- $W = (D^t, DVM^t, F^t, (N_{DV}{}^t))$:	Dokumentmenge, Dokumentvektorenmenge, Merkmalsmenge, (Klassifikation $N_{DV}{}^t$), (Wörterbücher, Thesauri u.ä. bleiben hier unberücksichtigt)
- $Q = Q(\Theta)$:	Menge aller Zeichensequenzen über einem endlichen Alphabet als Menge aller möglichen bzw. zugelassenen Queries.
- $I = ret(DVM^t, q_i, d_{DVR})$:	Retrievalfunktion als Funktion der Dokumentvektorenmenge, eines Queryvektors und der Metrik des Dokumentvektorraums DVR.
- $E \in \{P^{DVM(t)}, P^{D(t)}\}$:	Menge aller möglichen System-Outputs als Potenzmenge der Dokumentvektorenmenge oder Potenzmenge der Dokumentmenge.

1.2) Problemkomplexität des Information Retrievals

Das Ziel des Information Retrievals (IR, (Rijsbergen (1979[279]), Salton & McGill (1987[294]), Grossman & Frieder (1999[154])) ist die Präsentation einer Teilmenge oder geordneten Liste (Ranking) von Dokumenten aus einer Gesamt-Dokumentmenge als Output auf einen Input in Form einer Anfrage (Query) durch einen Nutzer (Agenten). Es wird angenommen, dass die Anfrage ein spezielles Informationsbedürfnis des Agenten repräsentiert, sodass die Ergebnis-Dokumentmenge zur Befriedigung dieses Informationsbedürfnisses beitragen soll. Die internen Selektionsoperationen des Information Retrieval Systems (IRS), welche eine Ergebnis-Dokumentmenge erzeugen, sollen derart sein, dass nach der Integration des Inhaltes der nachgewiesenen Dokumente in das kognitive System des Agenten, das Informationsbedürfnis gegenüber dem Zustand vor der Integration kleiner wird bzw. sich niveliert. Eine alternative Sicht geht nicht von der Verringerung des Informationsbedürfnisses aus, sondern allgemein von dessen Transformation in ein neues Informationsbedürfnis, d.h. der Constraint eines monoton fallenden Maßes an Informationsbedürfnis wird aufgegeben. Dadurch wird die Situation modellierbar, dass ein tiefer gehendes Verständnis eines Problemes andere oder sogar mehr Fragen aufwirft, sodass ein anderes bzw. größeres Informationsbedürfnis aus der Wissenserweiterung des Agenten durch Integration von Informationen aus nachgewiesenen Dokumenten folgen kann (siehe auch Swanson (1987[328])).

Das Ziel der IR-Forschung ist die Suche nach adäquaten und implementierbaren Modellen des IR-Prozesses, wobei Adäquatheit durch unterschiedliche Performancemaße wie z.B. Recall und Precision operationalisiert wird. In den letzten 40 Jahren IR-Forschung hat sich das Problem der Performanceverbesserung bezüglich der verwendeten Maßen als ausgesprochen hartnäckig dargestellt, was sich im Kontext des IR im Internet noch verschärft hat. Es stellt sich die grundsätzliche Frage, warum das IR ein so komplexes Problem ist, wofür die Kombination einer Reihe von Eigenschaften verantwortlich ist, die im Nachfolgenden skizziert werden sollen:

1) Hochdimensionale Zusammenhänge.
2) Nicht-lineare und multimodale Zusammenhänge.
3) Dynamische Zusammenhänge (nicht-stationäre Funktionen)
4) Quellen von Unsicherheit (uncertainty) und Vagheit (fuzzyness).
5) Diversität der Agenten und ihrer Ziele
6) Mehrziel-Anforderungen.

1.2.1) Hochdimensionale Zusammenhänge

Natürlichsprachliche Texte und insbesondere Fachtexte zeichnen sich durch eine große Anzahl darin enthaltener Terme aus, sodass zur Repräsentation dieser Texte ein entsprechend hochdimensionaler Parameterraum (Dokumentvektorenraum) notwendig ist. Dies muss kombiniert betrachtet werden mit einer großen Anzahl von Dokumenten, was im Fall vektorraum-basierter IR in einer hochdimensionalen Term-Dokument-Matrix mündet.

Der Versuch, durch unterschiedliche Verfahren eine Dimensionsreduktion dieser Räume zu erreichen, wurde dementsprechend in der IR-Forschung früh unternommen, wobei die Clusterung von Dokumenten und Termen im Zentrum der Bemühungen stand (Sparck Jones (1971[319]), Crouch (1972[84]), Robertson (1977a[283], b[284], c[285], 1979[286]), Salton (1971[293]), Panyr (1986a[247]), Bachelier (1995[14])). Neuere Verfahren versuchen mit einer Hauptkomponenten-Analyse (Principal Component Analysis (PCA); Oja (1993[240]), Diamantaras & Kung (1996[94]), Lee (1998[200])) eine Dimensionsreduktion zu erreichen wie das Latent Semantic Indexing (LSI; Dumais et al. (1988[100]), Furnas et al. (1988[136]), Bartell et al. (1992[27]), Soboroff et al. (1998[315]), Hofmann (1999[165]), Ando (2000[11])). Hierzu zählen auch konnektionistische Verfahren, die nicht-lineare Faktorenanalysen durch Self-Organizing-Maps (SOMs; Ritter et al. (1991[282]), Kohonen (1989[184], 1995[185]), Bachelier (1998a[15])) approximieren und dies auf den Fall des IR abstimmen (Scholtes (1993[301]), Bachelier (1995[14])).

In der Literatur zum Machine Learning und zur Optimierung ist die Problematik hochdimensionaler Zusammenhänge unter dem Begriff curse-of-dimensionality bekannt, der von Bellman (1967[41]) eingeführt wurde (siehe z.B. Venkatesh et al. (1992[348]), Warwick & Kárny (1997[357])). Bezeichnet wird damit das exponentielle Wachstum von Regionen eines Suchraumes bei einem linearen Wachstum der Dimensionsanzahl, womit der Suchaufwand in dem entsprechenden Raum bei deterministischen Verfahren ebenfalls exponentiell steigt. Als Lösungsansatz gegen den curse-of-dimensionality können Verfahren zur Randomization (siehe z.B. Traub et al. (1988[338]), Wozniakowski (1996a[372], b[373])) eingesetzt werden, wie z.B. Monte Carlo Verfahren (Fishman (1995[114])), die unabhängig von der Dimensionalität des zugrunde liegenden Raumes arbeiten. Erkauft wird diese Eigenschaft damit, dass die gefundenen Lösungen nur mit einer bestimmten Wahrscheinlichkeit in einem definierten Lösungsintervall liegen, d.h. dass die Lösung nur für den sogenannten average-case gelten (siehe auch Abschnitt 5.2.3)).

1.2.2) Nicht-lineare und multimodale Zusammenhänge

Indexierungs- und Retrievalfunktionen in vektorraumbasierten Modellen verwenden nicht-lineare, reelle Gewichtungen von Merkmalen und Dokumenten. Dies ergibt sich daraus, dass es sinnvoll ist, Gewichtungen auf ein reelles Intervall wie [0, 1] zu normieren, da die Interpretierbarkeit beliebig großer oder kleiner Gewichte verloren geht. Das gleiche gilt für Schätzungen der Relevanz durch ein IRS, (siehe Kapitel 4)) wobei reelle Relevanzschätzungen als nicht-lineare, multimodale Funktion der Distanzen von Dokumentvektoren ein Schwerpunkt der vorliegenden Arbeit bildet.

Unimodale Funktionen sind Funktionen mit genau einem Extremwert. Eine unimodale Relevanzfunktion würde bedeuten, dass an genau einem Punkt im Dokumentvektorenraum ein Relevanzmaximum existiert, wobei das Dokument mit dem am nächsten liegenden Dokumentvektor das relevanteste Dokument ist. Notwendig muss von einer monoton fallenden Relevanzfunktion um das Maximum einer unimodalen Funktion ausgegangen werden, d.h. je weiter ein Dokumentvektor vom Relevanzmaximum entfernt liegt, desto geringer wird dessen Relevanz bewertet. Die Modellierung des Retrievals mit einem Queryvektor und einer unimodalen Funktion ist die Standardsituation in den vektorraumbasierten Modellen, d.h. das Relevanzmaximum wird an der Stelle des Queryvektors bzw. des am nächsten liegenden Dokumentvektors angenommen, und allen anderen Dokumentvektoren wird explizit oder implizit ein Relevanzwert als monoton fallende Funktion der Distanz zu dem Maximum zugeordnet. Diese Vorgehensweise ist stark simplifizierend, da mit einer unimodalen Funktion alle Fälle nicht modellierbar werden, bei denen mehrere Cluster relevanter Dokumentvektoren existieren, zwischen denen nicht-relevante Dokumentvektoren liegen. Diese Fälle lassen sich nur mit multimodalen Funktionen modellieren, bei denen jeder Cluster relevanter Dokumentvektoren durch ein eigenes Maximum der Relevanzfunktion beschrieben werden kann. Diese Argumentation wird in Kapitel 4) verwendet, um die effektive Verwendung multimodaler, nicht-linearer Relevanz-Approximationsmodelle zu begründen.

Demgegenüber verwenden probabilistische Modelle nicht-lineare Verteilungsannahmen, welche die Indexierung und das Retrieval als nicht-lineare Prozesse kennzeichnet, jedoch werden nur unimodale Verteilungen verwendet, die nur durch eine Kombination im Sinne eines „Mixture Models“ (McLachlan & Basford (1987[214])) zu einer multimodalen Funktion aggregiert werden können.

Nicht-lineare und multimodale Zusammenhänge in hochdimensionalen Räumen zu modellieren, ist eine mathematisch anspruchsvolle Aufgabe und erfordert große Rechen- und Speicherressourcen, sodass effektive Verfahren im IR entsprechend aufwendig werden.

1.2.3) Dynamische Zusammenhänge (nicht-stationäre Funktionen)

Die offensichtliche Dynamik in IRS besteht in der Aufnahme neuer Dokumente und deren Integration in die gegebene Repräsentationsstruktur, die sich dadurch lokal bzw. global verändern kann. Insbesondere beim Vorliegen von Verfahren zur Dimensionsreduktion kann die Integration neuer Informationsobjekte gravierende Folgen besitzen, wobei eine dynamische Anpassung einer Clusterstruktur eine vergleichsweise problemlose Reorganisation darstellt (Crouch (1972[84])). Verändert sich die Anzahl der Objekte (Informationsobjekte oder Cluster) in einem Raum eines vektorraum-basierten IRS, so verändert sich die Dimension der Merkmals-Dokument-Matrix, und somit die Dimension des komplementären Raumes über sequentiell und rekursiv abhängige Merkmals- und Dokument-Graphen (siehe Abschnitt 3.8); siehe auch Bachelier (1995[14])). Dies erfordert globale Reorganisations-Operationen oder im Extremfall einen Neuaufbau der Strukturen in dem betreffenden Raum, sodass die Anzahl der dynamischen Anpassungen unter dem Constraint der gewünschten Aktualität minimiert werden sollte.

Durch die Aufnahme neuer Dokumente besteht immer die Chance, dass darin Terme enthalten sind, die bislang dem IRS unbekannt waren. Eine Aufnahme neuer Terme verändert die Dimension des Dokumen-

traumes, mit der Notwendigkeit der Reindexierung aller vorhandenen Dokumente und der Anpassung etwaig vorhandener Strukturierungen wie einer Klassifikation.

Diese dynamischen Prozesse der Reorganisation von Repräsentationsstrukturen machen die Retrieval-Funktion zu einer nicht-stationären Funktion, d.h. der gleiche Input kann zu unterschiedlichen Zeitpunkten einen unterschiedlichen Output erzeugen. Die nicht-stationäre Retrieval-Funktion ist von Ereignissen abhängig, die bezogen auf das IRS als extern zu bezeichnen sind, da das IRS keinen Einfluss auf die Produktion von Dokumenten und den darin enthaltenen Termen besitzt. Der einzige Einfluss besteht in der Entscheidung, ob neue Dokumente und Terme repräsentiert werden sollen. Ein IRS, das diese Entscheidung selbst treffen könnte, wäre autonom im Sinne autonomer Agenten, doch diese Entscheidung wird dem IRS ebenfalls von externen Instanzen in Form der Systembetreiber vorgegeben.

Es existiert ein Trade-of zwischen hochdimensionalen Zusammenhängen und dynamischen Funktionen beim vektorraum-basierten IRS in Bezug auf die Merkmalsanzahl. Werden keine Terme als Merkmale verwendet, sondern Zeichen-n-Gramme (Scholtes (1993[301]), Bachelier (1995[14])), so besitzt der Dokumentvektorenraum a priori die maximale Dimension, die somit nicht mehr verändert werden muss. Sei Θ ein endliches Alphabet mit $|\Theta|$ Zeichen (siehe Kapitel 3)), so besitzt ein Dokumentvektorenraum auf der Basis von Zeichen-n-Grammen eine Dimensionsanzahl von $|\Theta|^n$. D.h. jede Dimension des Dokumentvektorenraumes repräsentiert ein mögliches Zeichen-n-Gramm, wobei der Wert, der einer Komponente eines Dokumentvektors zugeordnet wird, d.h. die Merkmalsgewichtung, eine nicht-lineare Funktion der Häufigkeit des Auftretens des entsprechenden n-Gramms in dem zugrunde liegenden Dokument ist. Die Stabilität der Dimension des Dokumentvektorenraums wird somit erkauft durch seine große Anzahl an Dimensionen, wobei das Gesamtystem dennoch dynamisch bleibt, da Dokumente neu indexiert werden, was die Verteilungsstruktur im Dokumentvektorenraum, sowie die Dimension des Merkmalsraumes verändert.

Allgemein wird der Umgang, die Modellierung und Optimierung und Control von nicht-stationären Funktionen als schwierig bewertet, wobei robuste, adaptive Verfahren wie evolutionäre Algorithmen notwendig werden (Goldberg & Smith (1987[145]), Grefenstette (1992[153]), Dasgupta & McGregor (1992[89]), Cobb & Grefenstette (1993[69]), Vavak et al. (1996[346]), Mori et al. (1997[224]), Mori et al. (1998[225]), Stagge (1998[322]), Dozier (2000[97]), Santo & Kita (2000[295])). In Kombination mit der Nicht-Linearität dieser Funktionen und den hochdimensionalen Räumen trägt diese Systemeigenschaft entscheidend zu der Gesamtkomplexität bei.

Die dynamische Aufnahme neuer Informationsobjekte koppelt ein IRS an die offene Welt (Popper (1994[262], 1995[263]), Mühlenbein (1994[229], 1995[230])), was explizite Folgen für die Methoden hat, da wahrscheinlichkeitstheoretische Induktionsmethoden nur in geschlossenen Welten gelten (Popper (1994: 376ff[262])). Ein IRS als offenes System (dissipatives System; Nicolis & Prigogine (1977[235])) in einer offenen Welt zu betrachten, wurde jedoch in der IR-Forschung bislang nicht als Modellierungsansatz verwendet.

1.2.4) Unsicherheit (uncertainty) und Vagheit (fuzzyness)

In Motro & Smets (1997[228]) werden Ursachen und Vorschäge zum Umgang mit Unsicherheit und Vagheit in Informationssystemen allgemein und IRS im speziellen (Turtle & Croft (1997[343])) gemacht (siehe auch Natke & Ben-Haim (1997[232])). Als Ursachen werden von Turtle & Croft (1997:191f[343]) folgende Faktoren genannt:
a) Dokumentrepräsentation
b) Repräsentation der Informationsbedürfnisse
c) Retrievalfunktion
d) Repräsentation im Kontext des Relevanz-Feedbacks

Die Indexierungsfunktion als Vorgang der Dokumentrepräsentation in vektorraumbasierten IRS bildet ein Dokument bzw. ein Merkmalshäufigkeitsvektor auf einen Dokumentvektor als Element eines metrischen Vektorraumes ab. Die Auswahl der Merkmale definiert die Struktur der Repräsentation, sodass Unsicherheiten bezüglich einer adäquaten Auswahl bestehen. Werden aus einer Dokumentmenge Terme automatisch extrahiert, so muss spezifiziert werden, welche Zeichensequenzen als Terme verwendet werden sollen, wodurch eine andere Unsicherheit bezüglich der Adäquatheit besteht.

Neben der Merkmalsauswahl ist die Merkmalsgewichtung die wesentliche Quelle der Unsicherheit, da eine Vielzahl von Vorschlägen für Gewichtungsfunktionen gemacht wurden (siehe z.B. Singhal (1997[312]), Zobel & Moffat (1998[376])). Welche Indexierungsfunktion für eine gegebene Dokumentmenge in Abhängigkeit von bestimmten Eigenschaften, die den Agenten als zukünftigen Nutzern zugeschrieben werden, hinreichend gut ist, kann nicht vorhergesagt werden (Zobel & Moffat (1998[376])).

Die Formulierung des Informationsbedürfnisses des Agenten in Form einer Query, führt weitere Unsicherheiten, Vagheiten und subjektive Faktoren wie der aktive Gebrauch eines Vokabulars ein und wird in der IR-Forschung am häufigsten zitiert. Wird das IR im Kontext eines Problemlösungsprozesses betrachtet (siehe Abschnitt 3.9.1)), und wird eine neuro-kognitive Sichtweise vertreten, die von funktionalen Hirnarealen ausgeht, so bedeutet die Queryformulierung eine Projektion interner Repräsentationsstrukturen aus einem funktionalen Problemlösungsareal auf Strukturen eines Sprachsyntheseareals. Es kann dabei nicht einmal von einer Abbildung als linkstotale, rechtseindeutige Relation ausgegangen werden, d.h. dass alle momentan existierenden Repräsentationsstrukturen aus dem Problemlösungsareal auf eine Struktur im Sprachsyntheseareal abgebildet werden können. Es besteht somit die Möglichkeit, dass Repräsentationsstrukturen nicht abgebildet werden und dass abgebildete Repräsentationen nicht umkehrbar abgebildet werden können. Sind z.B. bestimmte Vokabularelemente dem Agenten unbekannt, so können kognitive Strukturen darauf nicht abgebildet werden, was eine interne Quelle der Vagheit darstellt.

Wird ein interaktives Retrieval wie ein Relevanz-Feedback betrachtet (siehe Abschnitt 3.9), siehe auch Panyr (1987b[252]), Harman (1992[157]), Buckley & Salton (1995[63]), Cool et al. (1996[77]), Dunlop (1997[101]), Lundquist et al. (1997[203])), d.h. werden mehrere Dokumentlisten durch das IRS nachgewiesen, die der Agent bewerten soll, so kommen subjektive Faktoren, bezogen auf das Verstehen der Dokumente hinzu. D.h. die Abbildung von internen Strukturen aus dem Sprachanalyseareal auf Strukturen des Problemlösungsareals bildet eine weitere Quelle von Unsicherheiten und Vagheiten.

Weiterhin berücksichtigt werden muss, dass die Relationen bzw. Abbildungen zeitlich variabel sind, d.h. in allen Arealen verändern sich sowohl die Strukturen der Repräsentionen als auch die Anzahl der Repräsentationen. Entsteht in einem Areal eine neue Repräsentation durch interne Prozesse, so muss es in einer anderen Region, in die Repräsentationen abgebildet werden, keine oder noch keine korrespondierenden Strukturen geben, auf die abgebildet werden könnte. Denkbar wäre ein Prozess, dass zunächst auf nicht korrespondierende oder Default-Strukturen abgebildet wird, und dass im Rahmen einer Konsistenzherstellung neue Strukturen gebildet werden, auf die künftig eine Abbildung der entsprechenden Repräsentation erfolgt.

Als dritter Punkt wird von Turtle & Croft (1997[343]) die Retrievalfunktion angeführt, die als ret(DVM^t, q_i^t, d_{DVR}, ε) definiert wurde als Funktion der Dokumentvektorenmenge, eines Queryvektors, der Metrik des DVRs und einem Distanzschwellenwert. Unsicherheit besteht bezogen auf die externe Festlegung der Indexierungsfunktion $A_{IR(Q)}$ der Query, der Metrik und des Schwellenwertes. Die Metrik ist zudem zentraler Punkt jeder Rankingfunktion, die eine Ergebnis-Dokumentvektorenmenge auf eine geordnete Dokumentvektorenliste abbildet.

Als vierter Punkt werden von Turtle & Croft (1997[343]) Repräsentationen im Kontext des Relevanz-Feedbacks erwähnt. Relevanz-Feedback-Verfahren machen ein IRS zu einem adaptiven IRS, wobei interne Repräsentationen des IRS kurzfristig bzw. langfristig modifiziert werden. Wie diese Repräsentationen modifiziert werden, wird durch Adaptions-Funktionen definiert, die extern festgelegt werden müssen, und somit die Quelle von Unsicherheiten bezüglich der adäquaten Modifikation darstellen. Z.B. gehen Queryvektor-Relevanz-Feedback-Verfahren davon aus, dass die Query-Indexierungsfunktion $A_{IR(Q)}$ nicht adäquat ist, sodass die Ursprungsquery mit Hilfe der in der ersten Iteration nachgewiesenen Dokumentvektoren und der Relevanzbewertungen durch den Agenten in einen neuen Queryvektor mit Hilfe einer Queryvektor-Adaptions-Funktion überführt wird. Die Verkettung von $A_{IR(Q)}$ und der Adaptionsfunktion wird implizit als neue und adäquatere Query-Indexierungsfunktion verwendet. Neben dem Queryvektor-Relevanz-Feedback (siehe Abschnitt 3.9.2)) wird die Veränderung der Dokumentvektoren (Dokumentvektor-Relevanz-Feedback (siehe Abschnitt 3.9.3)) verwendet, während die Modifikation anderer Repräsentationsformen in der IR-Literatur unüblich ist (siehe Abschnitte 3.9.4) bis 3.9.8)).

1.2.5) Diversität der Agenten und ihre Ziele

Im Abschnitt über dynamische Zusammenhänge wurde bereits die Integration neuer Dokumente und Merkmale als Form der Kopplung eines IRS mit der offenen Welt beschrieben. Da das IRS keinen Einfluss auf die Produktion von Dokumenten besitzt, ist dies ebenfalls unter dem Gesichtspunkt der Unsicherheit interpretierbar, d.h. die Entwicklung der Repräsentationen in einem IRS sind durch diesen externen Faktor mit Unsicherheiten verbunden. Als weiterer externer Faktor wird die Diversität der Agenten betrachtet, d.h. es ist unsicher, welche Agenten mit welchen kognitiven Strukturen mit dem IRS zukünftig interagieren werden und welche Ziele sie damit verfolgen bzw. welche Probleme sie damit lösen wollen, wenn das IR im Kontext eines Problemlösungsprozesses betrachtet wird. Es werden immer neue Themen, Interessen, Ziele und Probleme auf der Ebene von Individuen und Gruppen gebildet, sowie immer neue Nutzergruppen aus unterschiedlichen Kulturen mit den IRS interagieren.

Weiterhin entstehen immer neue Mikro-Kulturen (Tribes) in denen Individuen mit gleichen Interessen, Zielen und Problemen sich global zusammenschließen, wobei ein Individuum zu einer Vielzahl unterschiedlicher Gruppen gehören kann. Ein solches Individuum kann nicht mehr mit einem Nutzermodell beschrieben werden, sondern es müsste kontextabhängig, d.h. in Abhängigkeit von der Gruppe, zu der das Individuum sich zum Zeitpunkt der Interaktion zugehörig fühlt, Nutzermodelle erzeugt werden, was eine Form von Polymorphismus darstellt. Durch diese Veränderungen wird die Diversität der Nutzer nicht nur in den Kontext der Unsicherheit, sondern auch in den Kontext dynamischer Zusammenhänge, d.h. nicht-stationärer Funktionen gestellt. Eine Polyrepräsentation von Nutzermodellen wird somit ein immer wichtig werdender Faktor, zumal eine wachsende Geschwindigkeit, in der sich die Diversität der Themen, Interessen, Ziele und Probleme vergrößert, beobachtbar ist.

Weiterhin kann ein Nutzer in einem engen Zeitintervall verschiedene Anfragen stellen, die aus unterschiedlichen Zielen und Problemen resultieren, sodass auch die Bildung eines Nutzermodells durch das IRS keinerlei Garantien besitzt, um Unsicherheiten bezüglich der Ziele des Nutzers zu verringern. Das gleiche Argument kann für Langzeit-Nutzer und periodische Nutzer etwa im Kontext von Information-Filter-Systemen angewendet werden, deren Anfragen, Ziele und Probleme einer zeitlichen Drift unterliegen kann.

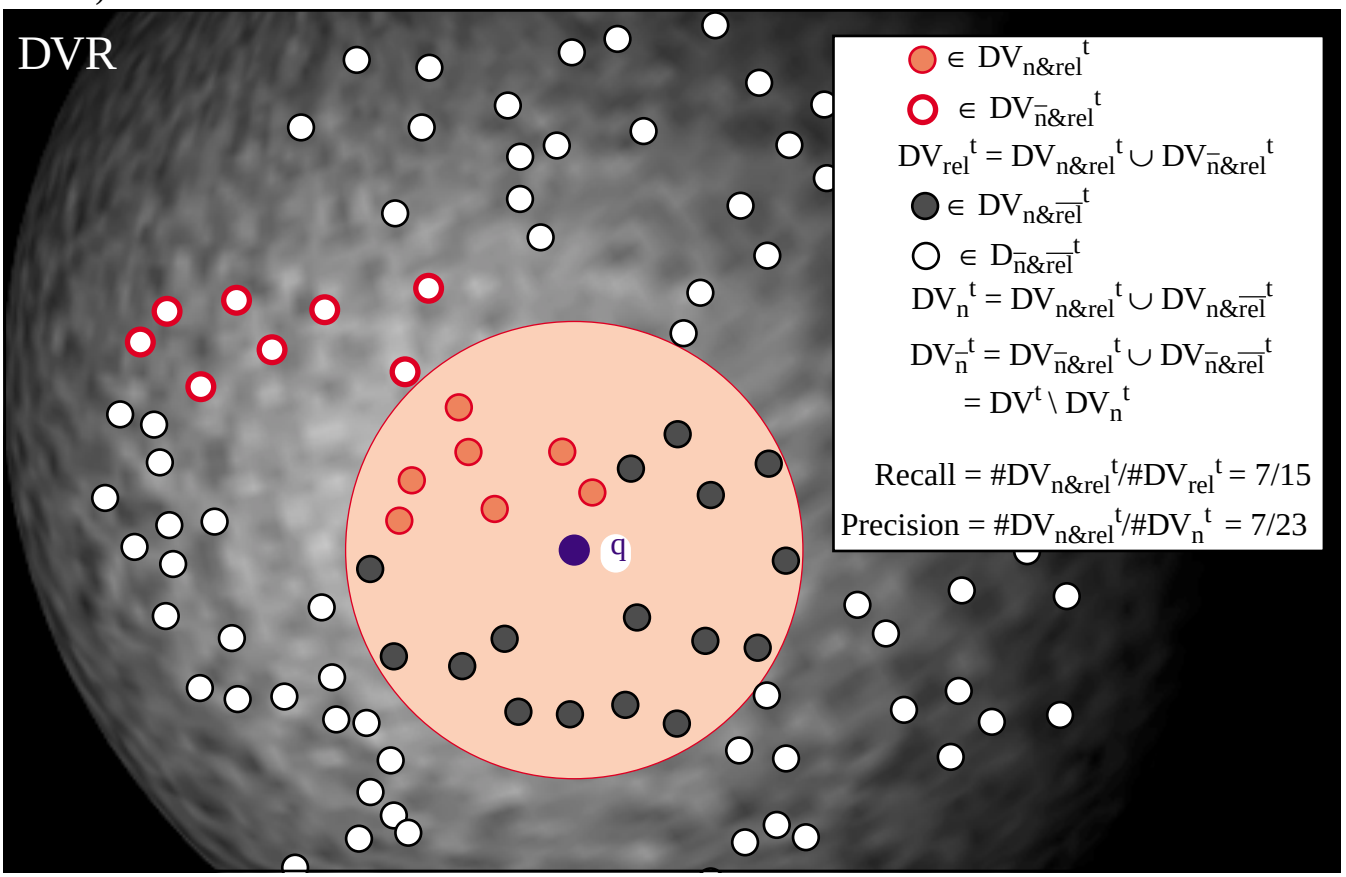
Die Bildung von Hypothesen über zukünftige Nutzer hat nicht nur Folgen für das Interfacedesign, sondern auch für die Wahl von Indexierungs- und Retrievalfunktionen. Der Versuch diese so zu wählen, dass einem „durchschnittlichen“ Agent geholfen wird, bestimmte Problemklassen zu lösen, ist eine sinnvolle Strategie, wenn von einer homogenen Agentenverteilung ausgegangen werden kann. Ist die Gesamtheit der Agenten jedoch bezüglich einer Vielzahl von Variablen divergent, und kann erwartet werden, dass das globale Leitbild der Individualisierung zu einer ständig größer werdenden Varianz in der Agentenverteilung führen wird, so ist die Auswahl einer einzelnen Indexierungs- und Retrievalfunktion und somit einer einzelnen Repräsentation von Informationsobjekten innerhalb des IRS keine geeignete Strategie. Diese Argumentation führt direkt zu dem Vorschlag der Polyrepräsentation in IRS (siehe Abschnitt 1.6) und Kapitel 3)).

1.2.6) Mehrziel-Anforderungen

Performancemessungen von IRS bezüglich einer Query werden durch Maße wie z.B. Recall als Quotient aus der Anzahl der nachgewiesenen relevanten Dokumente ($D_{n\&rel}^{t}$) und der Anzahl der gesamten relevanten Dokumente (D_{rel}^{t}) und durch Precision als Quotient aus der Anzahl der nachgewiesenen, relevanten Dokumente und der Anzahl der nachgewiesenen Dokumente (D_{n}^{t}) bestimmt. In Abb. 1) werden die Dokumentvektoren anstatt der korrespondierenden Dokumente verwendet, um die benötigten Mengen und die Definition von Recall und Precision zu illustrieren. Recall und Precision stellen komplementäre Maße dar, d.h. die Maximierung des einen Maßes minimiert das andere Maß. Die daraus ableitbare Generalisierung, dass die Optimierung eines IRS ein Mehr-Ziel-Optimierungs-Problem (Rosenthal (1985[290]), Steuer (1986[324]), Ringuest (1992[281]), Dasgupta et al. (1999[90]), Veldhuizen (1999[347]), Zitzler (1999[375])) darstellt, wurde in der IR-Literatur jedoch nicht explizit dargestellt.

Die Mehr-Ziel-Sichtweise gilt insbesondere, wenn über Recall und Precision hinaus zusätzliche Performancemaße verwendet werden, die jeweils bestimmte Charakteristika von IRS beschreiben, und die teilweise Überlappungen untereinander bezogen auf die gemessenen Charakteristika besitzen. Erwähnt werden kann z.B. die Utility als Differenz der gewichteten Anzahl der relevanten Dokumente und der gewichteten Anzahl der nicht-relevanten Dokumente (siehe Vogt (1999: 11[352])). Weiterhin können Performancemaße verwendet werden, die auf Rang-Korrelationen basieren (Bartell (1994[28])) und Maße, die sich aus der Signal-Detection-Theory (Egan (1975[107]), Swets (1996[329])) ableiten (siehe Vogt (1999:11f[352])).

Abb. 1) Definition von Recall und Precision



Ohne eine Menge von Performancemaßen ergibt sich eine Mehr-Ziel-Anforderung an IRS direkt aus den restlichen fünf beschriebenen Problemstellungen. D.h. das Handling hochdimensionaler, nicht-linearer und dynamischer Zusammenhänge unter dem Vorliegen von Unsicherheit und Vagheit in einem IRS ist bereits eine Mehr-Ziel-Anforderung.

1.3) Methodentransfer

Die vorgestellten Einzelprobleme und ihre Kombination sind nicht ausschließlich für das IR spezifisch, d.h. es existieren andere Untersuchungsbereiche, in denen ähnliche Problemfelder auftauchen. Z.B. zeichnet sich ein Produktionplanungssystem mit einer Menge unterschiedlicher Maschinen und einer zeitlich geordneten Liste an Jobs durch einen hochdimensionalen, diskreten Suchraum aus, in dem durch Schedulingverfahren nach optimalen bzw. suffizienten Maschinenbelegungsplänen gesucht werden muss

(Blazewicz et al. (1996[50])). Die damit verbundenen Probleme sind nicht-linear und nicht-stationär, wenn der Ausfall von Maschinen und der Ausfall bzw. die neue Aufnahme von Jobs disponiert werden müssen. Durch die Nicht-Beeinflussbarkeit dieser Faktoren durch das System ist es an die offene Welt gekoppelt, was Unsicherheiten und Vagheiten erzeugt. Durch zeitliche Constraints und monetäre Constraints entstehen zudem Mehr-Ziel-Anforderungen an die Leistung eines solchen Systems, sodass quasi alle aufgezeigten Problemfelder ebenfalls vorliegen.

Eine andere Domäne, in der Problemfelder dieser Art auftreten, ist die menschliche Kognition und ihre Modellierung durch kognitive Architekturen (Anderson (1990[8], 1991[9], 1993[10]), Newell (1990[233]), VanLehn (1991[345])). Dies dürfte nicht überraschen, da die Prozesse des IR, d.h. Indexierung und Retrieval, in Bibliotheken von Menschen lange vor der Verwendung von Computern durchgeführt wurden und immer noch durchgeführt werden. Wird das IR als Problem betrachtet, das von einem kognitiven System hinreichend gut gelöst werden kann, so liegt der Schluss nahe, kognitive Architekturen direkt als IRS zu verwenden. Dies ist jedoch mit einer Vielzahl von Problemen verbunden, die alle darin begründet sind, dass eine Menge von kognitiven Architekturen vorgeschlagen wurde, jedoch keine Architektur bezüglich Effektivität und Effizienz in den relevanten Problemstellungen dominierend ist. Doch selbst bei der Verwendung einer der Architekturen, die man bezogen auf Effektivitätskriterien als eine gute Architektur bezeichnen kann, ist ihre Leistungsfähigkeit gegenüber einer menschlichen Kognition immer noch bescheiden. Auf die kognitive Psychologie zu warten, um damit die Probleme des IR zu lösen, wäre eine vergleichbar verfehlte Strategie der IR-Forschung wie das Warten auf die KI in den 1980'er Jahren.

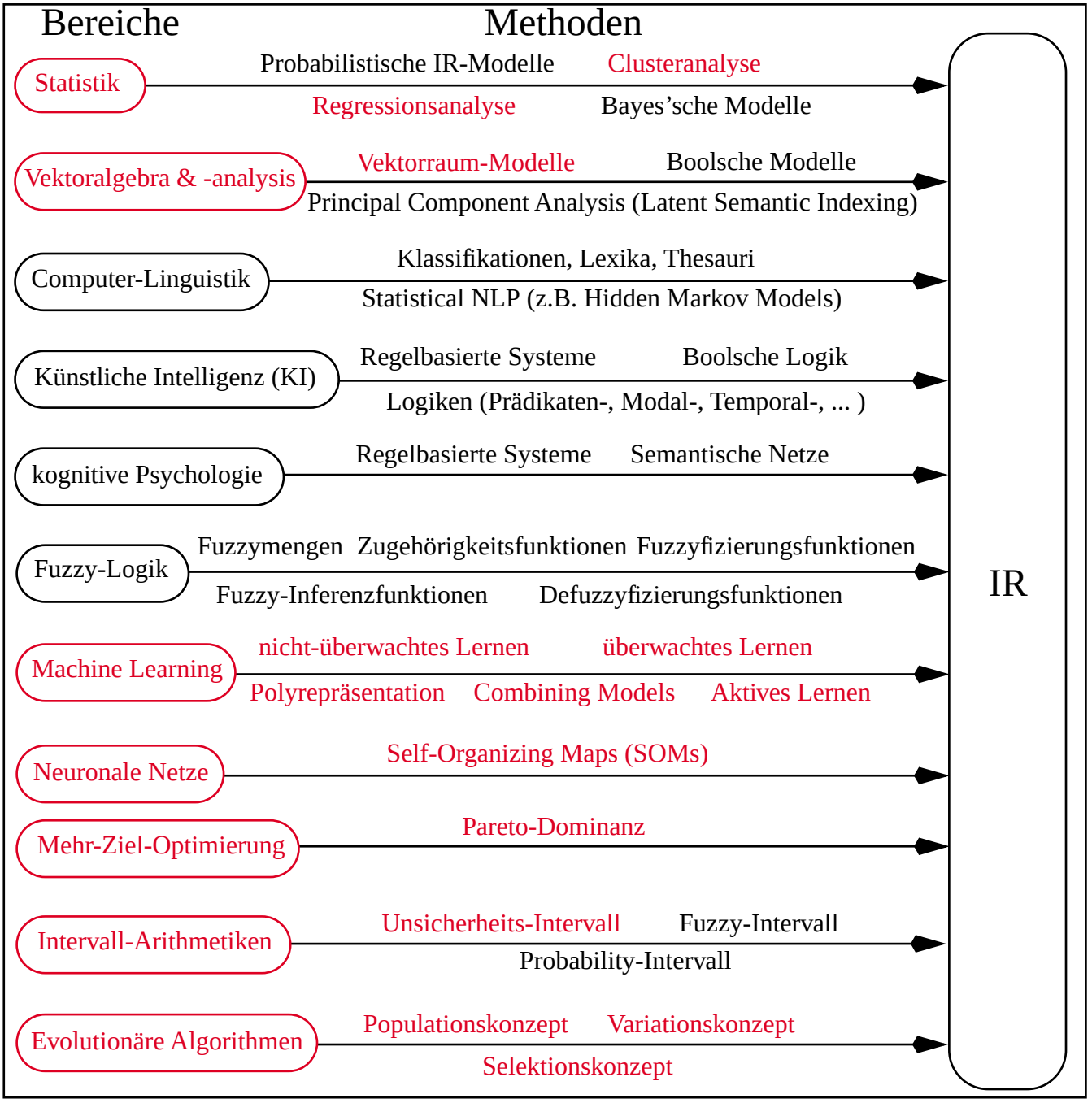
Es existiert eine Reihe weiterer Forschungsbereiche, welche mit gleichen bzw. ähnlichen Einzelproblemen konfrontiert sind, wie z.B. Machine Learning, Statistik, Optimierung und Control-Theory. In diesen Forschungsbereichen wurden eine Vielzahl von Methoden entwickelt, die zur Lösung der bereichsspezifischen Probleme eingesetzt werden. Durch die Erkenntnis, dass es sich um gleiche bzw. ähnliche Einzelprobleme handelt, wird ein Methodentransfer initiiert (siehe Abb. 2)), der über die gesamte Forschungshistorie des IR verfolgt werden kann. D.h. eine häufig beobachtbare Forschungslogik im IR besteht im Transfer von Lösungsmethoden aus Bereichen, in denen ähnliche Problemstellungen auftreten.

Betrachtet werden die Forschungsbereiche und die transferierten Methoden in Bezug auf die beschriebenen Problemfelder des IR, wobei Interdependenzen zwischen den Forschungsbereichen nicht detailliert beschrieben werden sollen wie z.B. Statistik und Computerlinguistik mit statistischen Natural-Language-Processing-Methoden (siehe jedoch Sprachlernen mit Evolutionären Algorithmen (Witten & Smith (1996[361]))).

Statistik ((Overbeck-Larisch & Dolejsky (1998[242]), Schuchmann & Sanns (1999a[303], b[304]), Abell & Braselton (1999[2])) beschreiben das Auftreten und die Eigenschaften von nicht-deterministischen Prozessen, d.h. es werden Unsicherheiten mit Hilfe von Konzepten wie z.B. Verteilungen bei der parametrischen Statistik beschrieben. Die direkte Methodenübertragung sind probabilistische Modelle des IR (Robertson (1977a[283], b[284], c[285], 1979[286]), Panyr (1986c[249]), Harter (1975[159]), Wong & Yao (1995[371]); Crestani et al. (1998[81])), bei denen z.B. für Merkmale die Wahrscheinlich-

keit berechnet werden, dass sie ein Dokument hinreichend gut beschreiben, oder bei denen für Dokumente die Wahrscheinlichkeit berechnet wird, dass sie bezogen auf eine Query relevant sind. Verbunden ist damit das „Probability Ranking Principle“ (Robertson (1977c[285])), bei dem die Dokumentergebnismenge in eine geordnete Liste überführt wird, indem als Ordnungsprinzip die fallende Wahrscheinlichkeit der Relevanz eines Dokumentes bezüglich der aktuellen Query verwendet wird.

Abb. 2) Methodentransfer ins Informations Retrieval



Die Zuordnung von Relevanzwahrscheinlichkeiten bei probabilistischen Modellen verwendet unterschiedliche Verfahren der schließenden Statistik wie Inferenz-Netze (Turtle & Croft (1990[341], 1991[342]), Turtle (1991[344]) oder Inferenzstrategien auf der Basis von Bayes-Verfahren (allg. siehe Martignon & Laskey (1995[211]); Turtle & Croft (1990[341], 1991[342]), Cooper (1991[78]), Kwok (1995[194]), Larkey & Croft (1996[196]), Henrion et al. (1997[161])).

Clusteranalysen (Eckes & Rossbach (1990[103]), Jain et al. (1999[175]), McLachlan & Basford (1987[214])) gehören in den Bereich der nicht-parametrischen Statistik (Schuchmann & Sanns (1999a[303])) und erzeugen aus einer Gesamtmenge von Objekten wie Dokumenten oder Merkmalen entsprechend der Repräsentation dieser Objekte Teilmengen (Cluster), die flach oder hierarchisch strukturiert werden können. Wie bereits beschrieben, werden Clustermethoden im IR im Kontext der hochdimensionalen Repräsentationen der Informationsobjekte eingesetzt (curse-of-dimensionality, Bellmann (1967), Warwick & Kárny (1997[357])).

Mit Regressionsmodellen (Ryan (1997[292]), Rawlings et al. (1998[273])) wird versucht, einen funktionalen Zusammenhang zwischen zwei oder mehr Variablen zu beschreiben, mit denen Objekte repräsentiert sind. Im Rahmen dieser Arbeit nimmt die Relevanz-Regression als Spezialfall der Relevanz-Approximation eine entscheidende Stellung ein (siehe Abschnitt 1.7) und Kapitel 4), siehe auch Cooper et al. (1992[79]); Gey (1994); Robertson & Walker (1994[287])). Die Zuordnung von Relevanzwerten zu Dokumenten bezüglich einer Query durch Regressionsmodelle ist eine alternative Vorgehensweise zu der Zuordnung von Relevanzwahrscheinlichkeiten in probabilistischen IR-Modellen, wobei Relevanzwerte bei Regressionsverfahren aus dem Intervall [0, 1] auch als Wahrscheinlichkeiten eines probabilistischen Modells interpretierbar sind.

Das Vektorraum-Modell des IR (Salton (1971[293]), Panyr (1987a[251])), auf das sich die vorliegende Arbeit bezieht, baut auf Begriffen wie Vektorraum, Distanzfunktion und Metrik auf, die aus der Vektoralgebra und Vektoranalysis (Strampp (1996[325]), Trostel (1993[339], 1997[340])) stammen. Verfahren des Latent-Semantic-Indexing (Dumais et al. (1988[100]), Furnas et al. (1988[136]), Bartell et al. (1992[27]), Soboroff et al. (1998[315]), Hofmann (1999[165]), Ando (2000[11])) basieren ebenso auf dem Vektorraummodell wie konnektionistische Verfahren des IR (Scholtes (1993[301]), Bachelier (1995[14]), Cunningham et al. (1997a[85], b[86])).

Boolsche Modelle des IR (Waller & Kraft (1979[354])) sind die ältesten Modellierungen, die Beziehungen zum Vektorraum-Modell sowie zu logischen IR-Modellen besitzen (Boolsche Logik). Grundlage von Queries sind Terme, die mit den Operatoren „Und“ ($\wedge$), „Oder“ ($\vee$), „Nicht“ ($\neg$) verknüpft werden, wobei Klammern hinzu kommen können, um strukturierte, hierarchische Ausdrücke zu erzeugen. Dokumente werden im einfachsten Fall als Binärvektoren repräsentiert, d.h. jedem Term wird ein Komponent des Binärvektors zugeordnet, wobei die Komponente mit 1 belegt wird, wenn das Merkmal in dem zu repräsentierenden Informationsobjekt auftritt bzw. die Komponente wird mit 0 belegt, wenn das Merkmal nicht auftritt. Eine Retrievalfunktion prüft, ob ein Dokument bezüglich der Query passend ist, d.h. ob bestimmte Terme enthalten sind oder ob bestimmte Terme nicht auftreten. Nachgewiesene Dokumente können nicht unterschieden werden, sodass als Ergebnis eine Menge und keine nach einem Rankingkriterium geordnete Liste vorliegt. Erweiterte Boolsche Modelle versuchen ein Ranking zu erzeugen, indem sie Gewichte der Terme in den Dokumenten bzw. in der Query verwenden, wenn diese berechnet werden bzw. vom Agenten spezifiziert werden (Lee (1994[198]), Grossman & Frieder (1999:58ff[154])). Dies kennzeichnet einen Übergang zu vektorraumbasierten Modellen, da eine Distanzfunktion zwischen einer Query und einem Dokument auf der Basis der Terme in der Query und den Gewichten als Distanz in einem metrischen Vektorraum betrachtet werden kann.

Computer-Linguistik allgemein versucht die Spracherkennung, das Sprachverstehen (Ram & Moorman (1999[270])) und die Spracherzeugung beim Menschen zu modellieren und zu implementieren (siehe z.B. Batori et al. (1989[29])), wobei eine Vielzahl symbolischer sowie subsymbolischer, d.h. konnektionistischer Ansätze (z.B. Scholtes (1993[301])) vorgeschlagen wurden. Zu den symbolischen Ansätzen gehören die Verwendung von Klassifikationen, Lexika und Thesauri, während statistische Methoden der Verarbeitung natürlicher Sprache (Statistical Natural Language Processing (Manning & Schütze (1999[210])) die Schnittmenge zwischen den symbolischen und konnektionistischen Ansätzen darstellen. Zu den statistischen Verfahren gehören insbesondere Hidden Markov Modelle (Buchholz (1991[60]), Sinclair (1992[311]), Davis (1993[92]), Buchholz et al. (1994[61]), Elliott et al. (1995[109])), für die auch Anwendungen in IR existieren (Miller et al. (1999[219])). Da in dieser Arbeit symbolische Ansätze wie die Verwendung von Lexika und Thesauri nicht betrachtet werden, besteht ein Anknüpfungspunkt bei den statistischen und konnektionistischen Verfahren, wobei insbesondere der Ansatz zu erwähnen ist, dass Dokumente, Queries und Merkmale (Features) als Zeichenketten eines endlichen Alphabetes betrachtet werden. Etwas periphär werden im Rahmen der Polyrepräsentation Sprachgenerierungsmodelle betrachtet, die aus einem Dokument oder einer Query eine Menge von Dokumenten bzw. Queries erzeugen, die nachfolgend indexiert werden. Dabei werden insbesondere Zeichenketten-Resampling-Verfahren betrachtet wie das Moving-Blocks-Verfahren (siehe Abschnitt 2.2.3)).

Künstliche Intelligenz (KI) (Görz (1993[151]), Bundy (1997[64])) versucht allgemein Hard- und Software zu entwickeln, die Leistungen erbringen, für welche beim Menschen das Attribut intelligent verwendet würde. Dabei wird keine biologische oder psychologische Plausibilität wie bei Neuronalen Netzen oder Modellen der kognitiven Psychologie gefordert, sondern es werden ausschließlich Effektivität und Effizienz als Performancemaße verwendet. Die KI hat seit Mitte der 50’er Jahre eine große Anzahl von Repräsentationsformen und Lernformen entwickelt, von denen einige im Kontext des IR eingesetzt wurden, wobei regelbasierte Systeme und logikbasierte Systeme angesprochen werden sollen.

Regelbasierte Systeme (Produktionssysteme; Newell (1990[233])) besitzen eine Wissensbasis in Form einer Menge von Wenn-Dann-Regeln. Kombimiert wird diese Wissensbasis mit einen Regelinterpreter, der in Abhängigkeit von einer Menge momentan aktiver Wissenselemente in einem Arbeitsspeicher (Working-Memory) ermittelt, welche Regeln anwendbar sind, indem er die Wissenselemente mit den Wenn-Teilen (If-Teile) der Regeln vergleicht (matched). Je nach der Art des Interpreters führt er alle anwendbaren Regeln aus, eine Teilmenge oder die erste passende Regel, die er findet. Der Dann-Teil (Then-Teile) der Regeln kann Wissenselemente im Arbeitsspeicher modifizieren, löschen, neue erzeugen, externe Aktionen auslösen u.a. Die Regeln in der Wissensbasis werden meist von Menschen erzeugt, bzw. es werden Lernmechanismen implementiert, die neue Regeln erzeugen können, was den Bereich des Machine Learnings berührt.

Der Einsatz von regelbasierten Systemen im IR überschneidet sich mit der Argumentation des Einsatzes von Modellen der kognitiven Psychologie im IR. Die Hypothese der kognitiven Psychologie, dass kognitive Prozesse durch Produktionssysteme adäquat modellierbar sind (Anderson (1990[8], 1991[9], 1993[10]), Newell (1990[233])), ist der Grund, warum diese Systeme im IR eingesetzt werden. Die Indexierung wurde und wird von menschlichen Experten durchgeführt, bevor Computersysteme sich an dieser Aufgabe versuchten. Ziel ist es in diesem Kontext, die Indexierungsregeln zu modellieren, von denen