Polyrepräsentation

Markus Wilmsen

Interaktives Information Retrieval - Seminar,
Sommersemester 2009

Informationssysteme – Prof. Dr.-Ing. Norbert Fuhr,
Dipl.-Inform. Thomas Beckers

Inhaltsverzeichnis

1. Vorwort..... 3

2. Vorüberlegungen

2.1. Kognitiver Ansatz im Information Retrieval..... 3

2.2. ASK – Hypothese..... 3

2.3. Ingwersens kognitives Modell..... 4

3. Polyrepräsentation

3.1. Definition..... 5

3.2. Polyrepräsentationsmodell im Information Retrieval 6

3.3. Beispiel..... 7

4. Studien

4.1. Evaluierungsmaßstab..... 9

4.2. Überblick..... 9

4.3. Mette Skov und Andere..... 10

5. Fazit und Ausblick..... 12

6. Literatur- und Quellenverzeichnis..... 13

1. Vorwort

Ziel dieser Arbeit ist die Definition und die Darstellung des Prinzips der Polyrepräsentation im *Interactive Information Retrieval* (*IIR*)-Prozess. Polyrepräsentation ist ein Modell aus dem Bereich des *Information Searching* und wurde von Peter Ingwersen auf der Basis seines kognitiven Modells (*Ingwersen Cognitive Model*) entwickelt und eingeführt. Durch Erweiterung der *ASK-Hypothese*, welche auch als Ausgangspunkt für sein kognitives Modell dient, versucht Peter Ingwersen unter Benutzung aller Informationen, sowohl aus dem kognitiven Bereich des Benutzers, dessen sozialem und/oder organisatorischem Umfeld als auch dessen Einfluß auf das Informationssystem, die Suchergebnisse zu verbessern. Wie innerhalb dieser Ausarbeitung anhand mehrerer Studien, insbesondere der von Mette Skov, gezeigt wird, führt die tatsächliche Implementierung von Polyrepräsentation zu deutlich besseren Ergebnissen beim *Interactive Information Retrieval*.

2. Vorüberlegungen

2.1. Kognitiver Ansatz im Information Retrieval

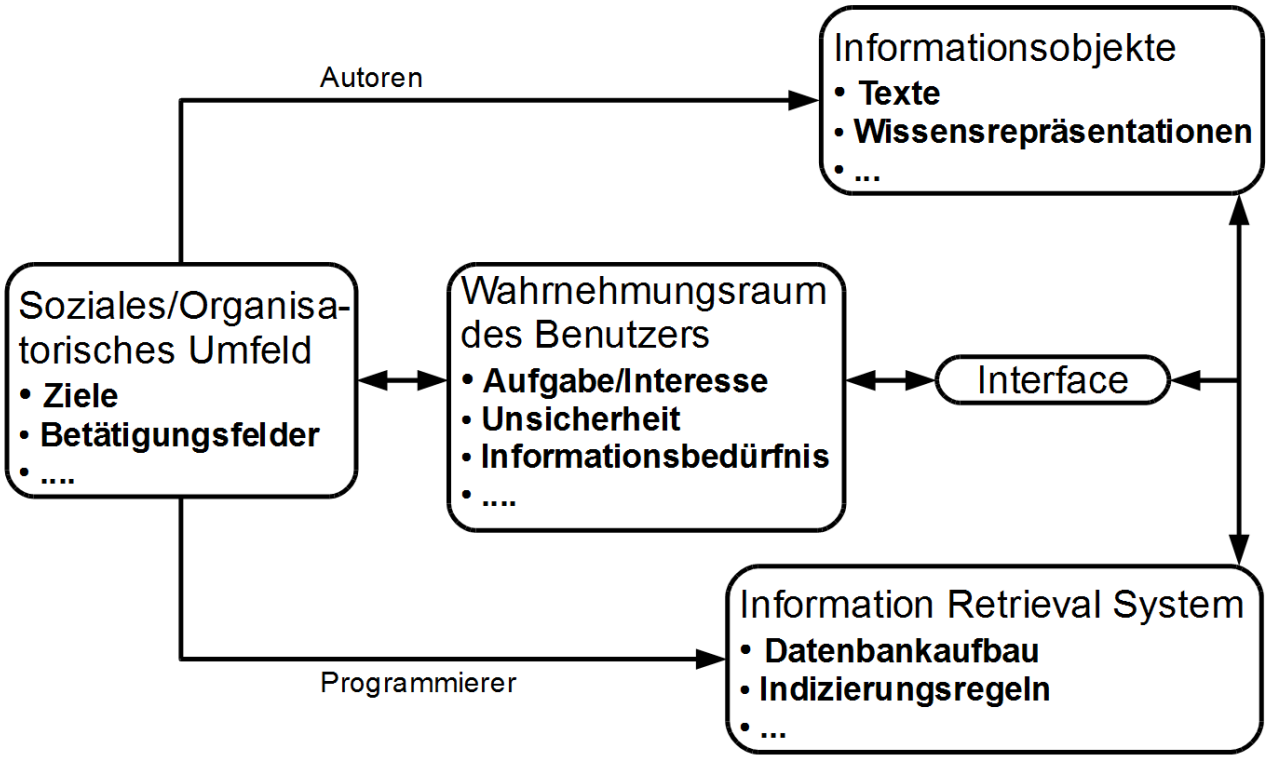
Die klassische Verarbeitung von Daten in einem Informationssystem findet immer auf der Ebene von linguistischen Symbolen statt. Zwar existiert hier ein Modell für die Vorgehensweise und deren Komponenten, wie z.B. Algorithmen, dieses wird aber dem System a priori, also bereits vor der Verarbeitung zum Erstellungszeitpunkt, implementiert. Die Verarbeitung von Informationen durch einen Menschen findet im sogenannten *Cognitive Space* des jeweiligen Menschen statt. Hierbei bezeichnet der *Cognitive Space* den Erkenntnis- bzw. Wahrnehmungsraum im Geist des Benutzers, also einen sehr dynamischen Komplex, der sich auch während der Verarbeitung von Informationen, also dem Erkenntnisprozess, verändern kann.
Das heißt ein Informationssystem ist bereits aufgrund seiner Architektur nicht in der Lage auf einem kognitiven Level Informationen zu verarbeiten und zu kommunizieren. Daraus ergibt sich als eigentliche Aufgabe für ein solches Informationssystem die Unterstützung des Benutzers bei der Suche nach Informationen, wobei es die dem *Information Retrieval (IR)* innewohnende *Unsicherheit* und *Vagheit* möglichst verringern soll. Das *IR* selbst muss vom jeweiligen Benutzer angegangen werden.[1,7]

2.2. ASK-Hypothese

Aus dem Verständnis heraus, dass ein *Information Need* ein spezieller Zustand im Geist des Menschen ist, der durch den Mangel von Informationen zu einem abnormalen Zustand im Wissen führt, und keineswegs eine klar definierte Frage, entwickelte Belkin[2] die sogenannte *Anomalous State of Knowledge-Hypothese*. Diese dem kognitiven Ansatz des *IR* folgende Hypothese versucht entsprechend die Informationen des Benutzers, die er auf der kognitiven Ebene kommuniziert, zu erfassen und fordert deren Berücksichtigung. Aufgrund der Unmöglichkeit der Formulierung des *Information Need* als konkreten *Request* oder gar *Query* sieht Belkin in der Kommunikation, z.B. von Benutzer und System, die einzige Alternative zur Lösung bzw. Aufhebung des *Anomalous State of Knowledge* und damit der Erfüllung des *Information Need*s. Ein Benutzer kann z.B. ein *Information Need* nicht formulieren, wenn ihm das enstprechende Vokabular fehlt, was insbesondere in der Frühphase des *Information-Retrieval-Prozess*es der Fall ist. **[1,2]**

2.3. Ingwersens kognitives Modell

Peter Ingwersen, ein Professor an der Royal School of Library and Information Science in Kopenhagen, der sich in diversen Arbeiten mit kognitiven Aspekten des *IR* beschäftigt hat, greift bei seiner Arbeit auch auf die *ASK - Hypothese* von Belkin zurück. Wie Belkin vertritt Ingwersen die These, dass der *Information-Retrieval-Prozess* interaktiv ist und dies Berücksichtigung finden muss, wenn die Ergebnisse eines *IR-System*s verbessert werden sollen. Bestärkt wird er hierbei sowohl durch von ihm aufgenommene Protokolle als auch durch die Arbeit anderer Informatiker wie Kuhltau, Murtonen und Järvelin.[1] Ingwersen erweitert allerdings die bisherige *ASK-Hypothese* um weitere kognitive Strukturen zu einem eigenen Modell, dem *Ingwersen Cognitive Model.*



Ingwersens kognitives Modell - Abb.1

Wie bei Belkin steht im Mittelpunkt die Interaktion von Benutzer bzw. dessen Wahrnehmungsraums mit dem *Interface* des *IR-System*s, welches wie üblich aus den a priori vorhandenen Modellen, Algorithmen und Daten besteht. Nach dem Modell von Ingwersen besteht der Wahrnehmungsraum hier aus verschiedenen Aspekten, wie dem aktuellen Wissensstand, der konkreten Aufgabe bzw. dem momentanen Interesse, der Unsicherheit und der aktuellen Version des *Information Need*. Als Erweiterung und damit Unterschied zu der *ASK-Hypothese* beinhaltet das *Cognitive Model* von Ingwersen noch die sozialen und organisatorischen Rahmenbedingungen des jeweiligen Benutzers. Abhängig vom Umfeld, bestehend aus seiner Domäne, also dem Bereich, aus dem es stammt, der Zielsetzung, der Aufgabe und den Vorstellungen und Vorlieben des Benutzers, findet durch Kommunikation mit diesem und dessen Wahrnehmungsraum eine beiderseitige Beeinflussung statt. Auf Seiten des Umfeldes geschieht dies in Form der Autoren des *IR-System*s und der *Informationsobjekte*. So hat die Kommunikation zwischen Benutzern und Autoren der *Informationsobjekte*, wie Texte oder Wissensrepräsentationen, Einfluss auf deren Gestaltung oder Löschung. Währenddessen nimmt der Benutzer durch den Autor kommunizierte Eigenschaften der *Informationsobjekte* auf, die seine zukünftigen *Information Need*s beeinflussen. Auf der anderen Seite führt Kommunikation zwischen

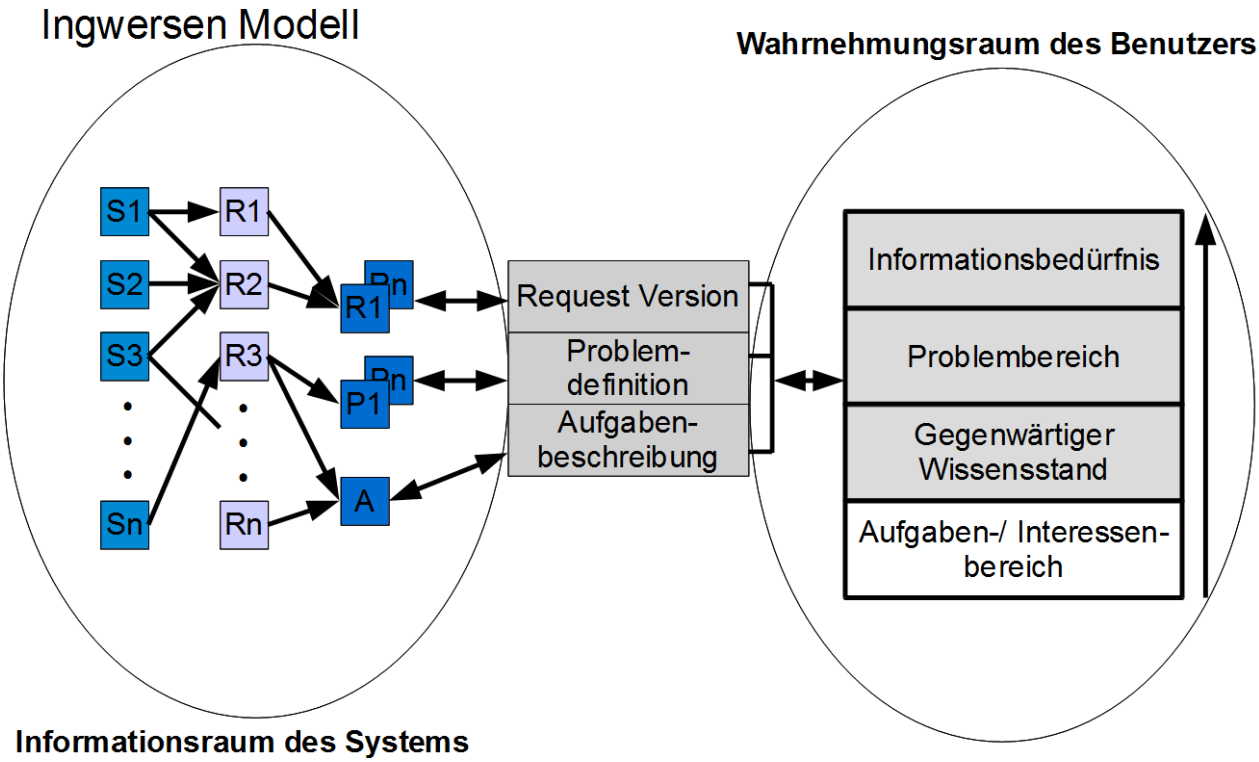
Benutzer und Autor des *IR-System*s zur Beeinflussung des Designs, sei es beim Layout der GUI oder den Möglichkeiten, die das *IR-System* anbietet. Umgekehrt wird die Benutzung des *IR-System*s durch den Benutzer durch Kommunikation mit dessen Autoren beeinflusst, indem z.B. Wege der Benutzung aufgezeigt werden.[1]

3. Polyrepräsentation
3.1. Definition

Basierend auf den Vorüberlegungen zur erweiterten *ASK-Hypothese* und seinem *Cognitive Model* entwickelt Ingwersen das Modell der Polyrepräsentation als *Intentional Redundancy*. Die immer vorhandenen, wenn auch nicht immer offen erkennbaren kognitiven Strukturen werden als potentiell wertvolle Informationen betrachtet und sollen möglichst umfangreich benutzt werden. Um dies zu realisieren, werden sogenannte *Overlaps*, also *Überschneidungen*, gebildet. Die Größe dieser *Überschneidungen* ist dann das Maß für die Relevanz des Dokuments für den aktuellen *Information-Retrieval-Prozess*. Hierbei werden verschiedene (*poly*, griechisch für *einige, mehrere*) Repräsentationen des gleichen *Informationsobjekt*s gebildet, was ein Dokument oder eine andere mit dem Dokument in Zusammenhang stehende Repräsentation, wie eine Liste von Schlüsselworten, sein kann. Unterschieden wird hierbei zwischen funktional und kognitiv unterschiedlichen Repräsentationen. Eine funktional verschiedene Repräsentation bezeichnet hierbei eine, die vom Autor der ursprünglichen Repräsentation erstellt wurde. Eine kognitiv verschiedene Repräsentation hingegen ist eine, welche evtl. vom selben Typ ist, aber nicht vom selben Autor stammt. So wären z.B. ein Blogeintrag, dessen Schlüsselwörter und Überschrift drei funktional unterschiedliche Repräsentationen, so sie alle vom selben Autor stammen. Ein professioneller Webkatalog, dessen kurze Texteinträge von Indexierern erstellt werden, besteht aus kognitiv verschiedenen Repräsentationen. Funktional unterschiedliche Repräsentationen erschließen also die Nutzung möglichst vieler kognitiver Strukturen aus dem Wahrnehmungsraum des Benutzers. Sie bedienen die *ASK-Hypothese.* Die kognitiv differenten Repräsentationen, die von Ingwersen in seiner Erweiterung der *ASK-Hypothese* und in seinem *Cognitive Model* als zu berücksichtigen angegeben wurden, erfassen die kognitiven Strukturen des sozialen und organisatorischen Umfeldes. Eine Besonderheit der Polyrepräsentation ist das mehrfache und somit redundante Speichern von Informationen. Dies steht dem normalen Vorgehen in der Informatik, nämlich möglichst performant zu arbeiten und Redundanz zu vermeiden, entgegen. Insbesondere bei kognitiv verschiedenen Repräsentationen kommt es zu einer, hier jedoch ausdrücklich gewünschten Redundanz, so dass das Konzept bzw. Modell der Polyrepräsentation auch als *Intentional Redundancy*, also beabsichtigte Redundanz, bezeichnet wird. Bildet also z.B. ein Indexierer für eine Sammlung von Dokumenten Schlüsselwörter und *Abstracts*, so ist zu erwarten, dass zentrale Begriffe im Dokument selbst, den Schlüsselwörtern und dem *Abstract* vorkommen. [1,4,8]

3.2. Polyrepräsentationsmodell im Information Retrieval

Um das Prinzip bzw. den Ablauf der Polyrepräsentation im *Interactive Information Retrieval* abzubilden gibt es von Ingwersen das *Global Model of Polyrepresentation in Information Retrieval*.[1]



Ingwersen Modell der Polyrepräsentation - Abb.2

Aufgeteilt ist das Modell aus Abbildung 2 in drei wesentliche Bereiche: *Information Space, Cognitive Space* und dem Interface des Informationssystems. Der *Cognitive Space* umfasst den Wahrnehmungsraum des Benutzers des Informationssystems.
Der *Cognitive Space* ist in vier Bereiche gegliedert, die, von unten nach oben gelesen, für die Entstehung und Gestalt des Informationsbedürfnisses verantwortlich sind. Das ist zum einen die Arbeitsaufgabe bzw. das Interessenfeld des Benutzers, das als statischer Rahmen am Anfang steht. Ist der Bildungsprozess des Informationsbedürfnisses angestoßen, wird dieses zunächst durch den aktuellen Erkenntnisstand weiter eingeschränkt. Erst jetzt, abhängig vom vorhandenen Wissen, erschließt sich der mit Unsicherheit belegte Problemraum für den Benutzer. Final ist dieser erst dann in der Lage das Informationsbedürfnis zu entwickeln. Hervorzuheben ist hier insbesondere der dynamische Charakter von aktuellem Erkenntnisstand, Problemraum und Informationsbedürfnis des Benutzers, da während eines Information-Retrieval-Prozesses die Wahrscheinlichkeit, dass sich diese Bereiche ändern, extrem hoch ist.
Die mehrfach erwähnte Dynamik des *Cognitive Space* bedingt nun außerdem, dass das Interface, also die Schnittstelle von *Cognitive Space* und *Information Space*, in der Lage sein muss, verschiedene Versionen von Benutzereingaben zu erfassen und zu verarbeiten. Dies ist hier durch verschiedene Schnittstellen im System (R, P, A) realisiert. Damit kommt dem Interface zur Implementierung und Nutzung der Polyrepräsentation eine zentrale Rolle zu.

Das System selbst, also der *Information Space,* besteht neben dem bereits erwähnten Interfaceschnittstellen aus zwei Komponenten:

1. Die *Semantic Entities* (S)*,* also die einzelnen, reinen Informationen
2. Die Repräsentationen bzw. Methoden von Repräsentationen (R), wie z.B. Thesauri, Schlüsselwörter und so weiter.

Hierbei bilden sich die Repräsentationen aus Kombination von *Semantic Entities* und damit zunächst nur aus Elementen des *Information Space*. Das Prinzip der Polyrepräsentation fordert explizit auch die Berücksichtigung der Informationen des *Cognitive Space* des Benutzers, daher werden nun im Bereich des Interfaces *Overlaps* aus den vorhandenen Repräsentationen des *Information Space* und den Informationen des *Cognitive Space* gebildet. Dies unterstreicht noch einmal die signifikante Rolle des Interfaces bei der Implementierung des Prinzips der Polyrepräsentation.[1]

3.3. Beispiel

Zur Illustration der Bildung von *Overlaps*, die sowohl Repräsentationen aus dem *Information Space* als auch aus dem *Cognitive Space* enthalten, und der Darstellung einer weiteren Folge der Implementierung der Polyrepräsentation, soll auf Basis der folgenden beschriebenen Dokumente und Repräsentationen ein Beispiel entwickelt werden.

1) Dokument zum Thema Auto

- Schlüsselwörter

Auto, Automobil, Mini

- Überschrift

Viel Ärger mit dem Mini. Viele kleine Mängel verderben den Spaß.

- Dokument

Klare Sache, der Mini ist ein Fall für Fans. Und davon gibt es reichlich, was auch die starke Nachfrage als Gebrauchtwagen erklärt. Wer sich für diesen Wagen entscheidet, sollte eine finanzielle Reparatur-Reserve anlegen.

2) Dokument zum Thema Kleidung

- Schlüsselwörter

Mini, Minirock

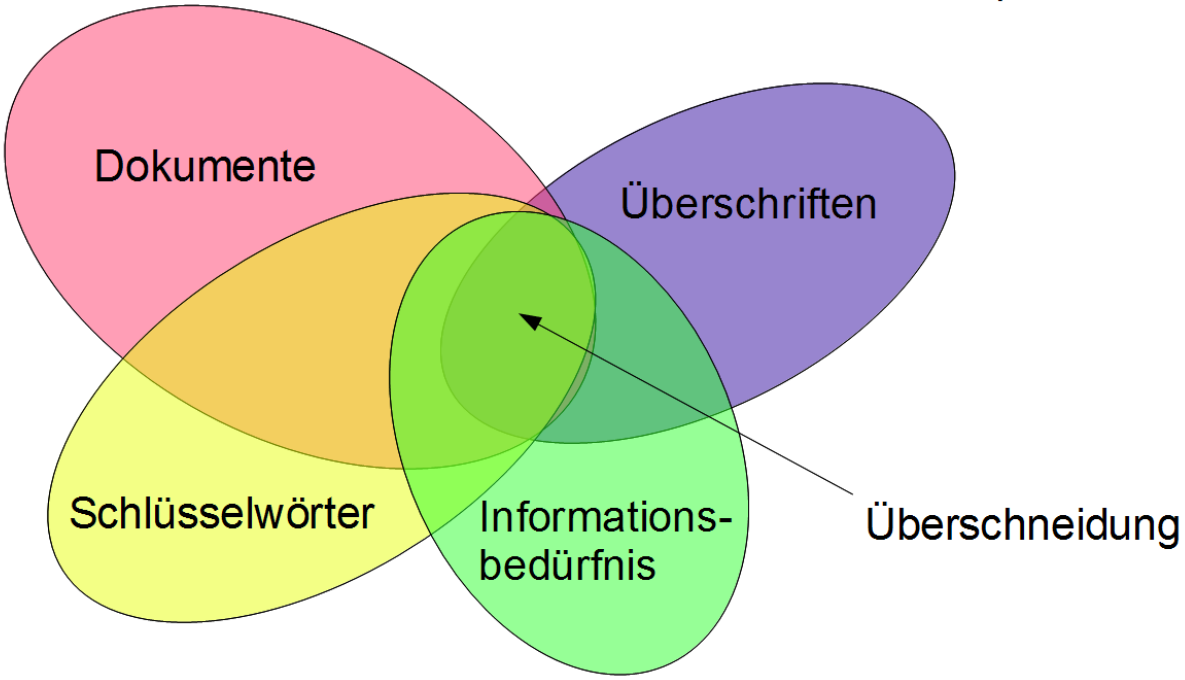
- Überschrift

Minirock

- Dokument

Der Mini ist ein Bekleidungsstück, das insbesondere die Mode der sechziger und siebziger Jahre geprägt hat. Es handelt sich um einen sehr kurzen Rock, der mindestens 10 cm über dem Knie der Trägerin endet.

Wie der Liste zu entnehmen ist, bestehen die Repräsentationen für die Dokumente aus dem *Information Space* aus dem Dokument selbst, der Überschrift und den Schlüsselwörtern. Es handelt sich in dem Fall um funktional unterschiedliche Repräsentationen, die kognitiv nicht verschieden sind, da sie vom selben Autor stammen.



Bildung von Überschneidungen für Polyrepräsentation - Abb.3

Gemäß Abbildung 3 wird bei der Bildung der *Überschneidungen* nun neben den Repräsentationen aus dem *Information Space* das Informationsbedürfnis als Repräsentation des *Cognitive Space* hinzugenommen. Das Interface, als wichtiges Element, soll in diesem Beispiel der Einfachheit halber aus einer Eingabe bestehen, die mit Hilfe von invertierten Listen und *boole'schem Retrieval* arbeitet.
Als zunächst erstes Informationsbedürfnis wird nun der Begriff „Mini" angenommen. Gemäß dem Modell von Polyrepräsentation nach Ingwersen werden nun also innerhalb des Informationssystems zunächst die Repräsentationen des *Information Space*, nämlich Dokument, Überschrift und Schlüsselwörter zusammen mit der Repräsentation des *Cognitive Space*, also dem Informationsbedürfnis, zu einer *Überschneidung* zusammengefasst. Das im Beispiel benutzte *bool'sche Retrieval* ergibt nun für das erste Dokument, das dem Themenkomplex Auto zuzuordnen ist, eine Überschneidung der Kardinalität vier, da alle Repräsentationen, also sowohl des *Information Space* als auch des *Cognitive Space* enthalten sind. Die Kardinalität der Überschneidung des zweiten, zum Themenkomplex Kleidung gehörenden Dokuments beträgt drei, da der Schnitt mit der Repräsentation Überschrift leer ist.
Aufgrund der bereits dargelegten Dynamik in den Bereichen des *Cognitive Space* ändert sich nun der aktuelle Erkenntnisstand und als unmittelbare Folge davon das Informationsbedürfnis. Für den nächsten Schritt im *Information-Retrieval-Prozess* lautet das Informationsbedürfnis nun „Minirock". Erneut werden die *Überschneidungen* aus *Information Space* und *Cognitive Space* gebildet. Durch das geänderte Informationsbedürfnis ändern sich die Kardinalitäten der *Überschneidungen* ebenfalls, so ist die *Überschneidung* für das erste Dokument auf Seite des *Information Space* leer. Die Kardinalität für die zweite Überschneidung ist gleich, allerdings enthält sie andere Repräsentationen. Somit ergibt sich als relevantes Dokument das zum Thema Kleidung gehörende.
Auffällig an der *Überschneidung* im zweiten Schritt bei dem Dokument zum Thema Kleidung ist, dass ein Dokument als relevant befunden wird, das den Begriff aus dem Informationsbedürfnis selbst nicht enthält und somit bei boole'schem Retrieval nur auf Basis der Dokumente gar nicht gefunden worden wäre. Dies impliziert bereits, dass bei Anwendung von Polyrepräsentation, die Menge der relevanten Dokumente besser abgedeckt wird.

4. Studien

4.1. Evaluierungsmaßstab

Zentrale Aussage von Ingwersen ist, dass durch Polyrepräsentation die Ergebnisse des *Information-Retrieval-Prozesse*s verbessert werden. Um dies bei Studien zu evaluieren, ist ein Maß für die Qualität der Suchergebnisse nötig. Dieses Maß ist in der Informatik in dem Bereich des *IR* die sogenannte *Precision*. Dieses Maß ist definiert als

$$\text{Precision} = \frac{|\text{gefundene Dokumente} \cap \text{relevante Dokumente}|}{|\text{gefundene Dokumente}|}$$

also Anzahl der relevanten Dokumente, innerhalb der gefundenen Dokumente im Verhältnis zur Gesamtzahl der gefundenen Dokumente. [9]

4.2. Überblick

Polyrepräsentation als solches ist bisher als Modell beschrieben worden, also als eine theoretische Überlegung aus dem Bereich Informatik. Zur Prüfung der Theorie und zur Evaluierung der Systeme mit Polyrepräsentation existieren inzwischen einige empirische Studien, die teilweise auch vom Lehrstuhl Peter Ingwersens selbst stammen. Auch wenn die bisherige Zahl der durchgeführten Studien noch nicht sehr umfangreich ist, so zeigt sich bei Art und Durchführung der existierenden Studien, dass jeder Aspekt der Polyrepräsentation betrachtet wird.
Exemplarisch für den Aspekt der Polyrepräsentation im *Cognitive Space*, also im Wahrnehmungsraum des Benutzers, hat Kelly zusammen mit anderen Wissenschaftlern eine Studie durchgeführt, bei der verschiedene Anfragen, also Formulierungen des Informationsbedürfnisses, desselben Informationssystembenutzers kombiniert werden. Lund wiederum hat zum Aufzeigen verschiedener Repräsentationen von Informationssystemeinstellungen die Ergebnisse von den zwölf besten *TREC5* Informationssystemen bzw. deren Kombination evaluiert. Hierbei ist die TREC (*Technical Retrieval Conference*) eine jährlich stattfindende Konferenz zum Thema *IR*, die als einen ihrer Hauptsponsoren die NIST (*National Institute of Standards and Technology*) verzeichnet[3]. Auch im großen Bereich des *Information Space* wurde Polyrepräsentation evaluiert. Hier ist die Studie von Larsen zu nennen, der verschiedene Repräsentationen von und aus Dokumenten heraus, auf Grundlage der *INEX* (*Initiative for the Evaluation of XML*)-Sammlung und des *INSPEC* (*Information Services in Physics, Electronics and Computing*)-Thesaurus' gebildet hat. Ganz ähnlich hat Mette Skov, eine Mitarbeiterin an Ingwersens Lehrstuhl, eine Studie im Bereich des *Information Space* durchgeführt. Allen Studien gemeinsam ist das Ergebnis, welches Ingwersen in seinen Annahmen, also Polyrepräsentation als Möglichkeit zur Verbesserung der Suche, bestätigt. Insbesondere zeigt die im folgenden Abschnitt ausführlicher behandelte Studie von Mette Skov, dass die Implementierung von Polyrepräsentation eine signifikante Verbesserung der *Precision* zur Folge hat. [4,5,6]

4.3. Mette Skov und Andere

Wie bereits erwähnt, gibt es von Mette Skov eine Studie zum Thema Polyrepräsentation im *Information Space.* An der aus dem Jahre 2004 stammenden Studie, die Mette Skov am Lehrstuhl von Ingwersen durchgeführt hat, soll im Folgenden insbesondere aufgezeigt werden, wie sich Polyrepräsentation auf die *Precision* auswirkt.
Rahmenbedingungen für die Studie bildet auf Systemseite ein *best match*-System, das das *probabilistische InQuery* benutzt, welches aus der Informatikabteilung der Universität von Massachusetts stammt. Als Daten- bzw. Dokumentenbasis dient die Mucoviszidose-Testdatenbank, die durch die Medline, die US-amerikanischen Nationalbibliothek für Medizin, gebildet wird. Aufgrund der relativ beherrschbaren Größe von 1.239 Dokumenten eignet sich diese Sammlung sowohl zur genauen Evaluierung von *Precision*, also dem Maß für die Relevanz der gefundenen Dokumente, als auch von *Recall*, also dem Maß der vorhandenen relevanten Dokumente, die auch gefunden wurden. Zudem ist es möglich, eine Gliederung in sehr und normal relevante Dokumente vorzunehmen. Die einzelnen Repräsentationen der Dokumente bestehen aus: Titel und Abstract, zusammengefasst in einer Repräsentation, den Referenzen auf andere Dokumente, beide vom Autor des Originaldokuments und damit funktional unterschiedlich, und sogenannten *MeSH*. Die mit *MeSH* abgekürzten *Medical Subject Header* bestehen ihrerseits aus den für die Suche benutzten Hauptthesen und Nebenthesen. Da sie von medizinisch geschulten Indexierern erstellt werden und nicht vom Autor des eigentlichen Dokuments, sind sie kognitiv unterschiedliche Repräsentationen. Abschließend ist noch festzuhalten, dass sowohl Anfragen in natürlicher Sprache als auch in stark strukturierter Sprache gestellt wurden.
Die wie angekündigt reduzierten Ergebnisse für die *Precision* können der folgenden Grafik entnommen werden.

	natürliche Sprache	stark strukturierte Anfrage	
Überschneidung/ Overlap	Precision relevant	Precision relevant	Precision stark relevant
Ol1 (Ti/Ab, Ht, Nt, Re)	41,00%	69,00%	53,00%
Ol2 (Ti/Ab, Ht, Nt)	13,00%	42,00%	20,00%
Ol3 (Ti/Ab, Ht, Re)	48,00%	79,00%	45,00%
Ol4 (Ti/Ab, Nt, Re)	29,00%	62,00%	47,00%
Ol5 (Ht, Nt, Re)	0,00%	64,00%	45,00%
Ol6 (Ti/Ab, Ht)	12,00%	45,00%	22,00%
Ol7 (Ti/Ab, Nt)	9,00%	27,00%	13,00%
Ol8 (Ti/Ab, Re)	9,00%	27,00%	19,00%
Ol9 (Ht, Nt)	6,00%	26,00%	14,00%
Ol10 (Ht, Re)	33,00%	38,00%	19,00%
Ol11 (Nt, Re)	21,00%	34,00%	16,00%
Ol12 (Ti/Ab)	2,00%	12,00%	55,00%
Ol13 (Hauptthesen)	10,00%	27,00%	12,00%
Ol14 (Nebenthesen)	4,00%	17,00%	7,00%
Ol15 (Referenzen)	5,00%	6,00%	2,00%

Ti/Ab – Titel und Abstract, Ht – Hauptthesen, Nt – Nebenthesen, Re – Referenzen

Ergebnisse der Studie Mette Skovs zur Polyrepräsentation (Auszug) - Abb.4

Zunächst sieht man, dass die These von Ingwersen, die *Precision sei* etwa proportional zur Größe der *Überschneidungen*, im Wesentlichen bestätigt wird. Betrachtet man die Anfragen in natürlicher Sprache, bei denen keine Unterscheidung von hoch und normal relevanten Dokumenten vorgenommen wird, zeigt sich, dass *Overlap 1*, also die erste *Überschneidung,* bestehend aus den Repräsentationen Abstract und Titel, Hauptthesen, Nebenthesen so wie den Referenzen, einen Wert für *Precision* von bereits 41% erreicht. Dieser Wert nimmt mit zunehmender Verkleinerung der Anzahl von Repräsentationen, mit denen die *Überschneidungen* gebildet werden, ab. So fällt die *Precision* für *Overlap 6*, die aus den Repräsentationen Titel und Abstract so wie den Hauptthesen besteht, auf lediglich 9%. Bei der Bildung von *Overlap 12*, bestehend nur noch aus der Repräsentation Titel und Abstract, erreicht der Wert für die Precision dann 2%, was bedeutet, dass von einhundert zurückgegebenen Suchergebnissen lediglich zwei für den Benutzer überhaupt relevant sind. Da bei den Anfragen in natürlicher Sprache außerdem keine Unterteilung zwischen normal und sehr relevanten Ergebnissen gemacht werden kann, ist dieser Wert als entsprechend schlecht zu bewerten.
Die Betrachtung der Werte für die *Precision* bei Anfragen in stark strukturierter Sprache zeigt das gleiche Ergebnis. Auch bei den strukturierten Anfragen ist *Overlap 1* mit einer Precision von 69% bei den normal und sehr relevanten Ergebnissen bzw. 53% bei den sehr relevanten deutlich besser als *Overlap 6*, die bei den normal und sehr relevanten Ergebnissen noch eine *Precision* von 45% so wie bei den sehr relevanten von 22% erreicht. *Overlap 12*, die mit der geringsten Anzahl von Repräsentationen gebildete *Überschneidung,* erreicht auch bei den strukturierten Anfragen ein deutlich schlechteres Ergebnis als die zwei Vorgänger, nämlich bei den normal und sehr relevanten noch eine *Precision* von 5% und bei den sehr relevanten von 8%.
Zwar bestätigen die Ergebnisse die prinzipielle Aussage der Korrelation von der Anzahl an Repräsentationen in einer *Überschneidung* und *Precision*, aber insbesondere der Unterschied von *Overlap 1* und dem kleineren *Overlap 3*, der einen höheren Wert für *Precision* ausweist, scheint gegen die prinzipielle Aussage zu sprechen, zumal der Unterschied sowohl bei natürlich sprachlichen als auch strukturierten Anfragen auftritt. Eine Betrachtung der beteiligten Repräsentationen und Hinzunehmen von *Overlap 4* zeigt, dass die Nebenthesen offensichtlich die Ergebnisse verschlechtern. Hieraus zieht Mette Skov den Schluss, dass eine Kombination von sowohl funktional als auch kognitiv verschiedenen Repräsentationen vermutlich nicht nur unnötig ist, sondern die Ergebnisse sogar verschlechtern kann und daher die Richtigkeit des Polyrepräsentationsmodells als solches nicht tangiert wird.
Ein weiteres wichtiges Ergebnis der Studie ist ein signifikant besseres Abschneiden von stark strukturierten Anfragen. Legt man den *Overlap 3* als Spitzenreiter bei den Werten für *Precision* an, so kann man eine Steigerung von 31 Prozentpunkten gegenüber einer natürlich sprachlichen Anfrage feststellen. Dieses Ergebnis lässt sich bei allen untersuchten Überschneidungen feststellen und impliziert daher für die Implementierung des Polyrepräsentationsmodells in ein *IR-System* die Nutzung von stark strukturierten Anfragen, um die Möglichkeit der *Precision*-Steigerung, die dem Modell innewohnt, wirklich voll nutzen zu können.[6]

5. Fazit und Ausblick

Als Fazit lässt sich zunächst erst einmal festhalten, dass Polyrepräsentation ein probates, also erprobtes Mittel zur Steigerung von *Precision* und Verringerung der *Unsicherheit* im *Information-Retrieval-Prozess* darstellt und damit die Verbesserung der Suche. Dies belegen die in dieser Ausarbeitung zitierten Studien bzw. die hier aufgearbeitete Studie von Mette Skov.

Natürlich gibt es auch im Polyrepräsentationsmodell Probleme. Die Bildung von Überschneidungen aus Repräsentationen benötigt eben genau diese. Dies kann auf verschiedene Weise geschehen. Analog zur Definition von funktional oder kognitiv unterschiedlichen Repräsentationen können als Quelle vom Autor des Dokuments oder von professionellen Indexierern erstellte Repräsentationen eingesetzt werden. Die Nutzung von funktional verschiedenen Repräsentationen beinhaltet die relativ große Gefahr von Manipulation durch den Autor. So hat die Nutzung von Schlüsselwörtern aus den Metadaten von Internetseiten dazu geführt hat, dass Betreiber einiger Seiten zum Erreichen besserer Platzierungen in Suchmaschinen teils völlig mit ihrer Seite unkorrelierte Schlüsselwörter eingesetzt haben. Selbst ohne gezielte Manipulation ist fraglich, wie groß die Gefahr ist, dass Autoren sich dazu hinreißen lassen, auch Themen, die in ihrer Arbeit vielleicht nur gestreift werden, bei den Schlüsselwörtern so zu listen, dass dies die Precision verschlechtert. Bei kognitiv differierenden Repräsentationen, besteht im Falle freiwilliger Autoren, wie sie zumindest in Ansätzen bei Wikipedia arbeiten, die Gefahr, dass es hier zu den gleichen Problemen wie bei den funktional abweichenden Repräsentationen von manipulierten und/oder überzogenen Repräsentationen kommt. Bei professionellen und damit bezahlten Autoren müssen Kosten und Nutzen abgewogen werden. Auch wenn bei ihnen eine Manipulation oder Ausschmückung eher unwahrscheinlich ist, kann insbesondere letzteres auch nicht ausgeschlossen werden, besonders bei Missinterpretation der Intentionen des Autors. Neben diesen Problemen im Bereich des *Information Space* gibt es auch im *Cognitive Space* offene Fragen. Zur tatsächlichen Implementierung von Polyrepräsentation im *Information Space* muss der Benutzer bereit sein persönlichen Informationen zur Verfügung stellen. Aufgrund von sogenannten *Sozialen Netzwerken* des *WEB 2.0* und dem teilweise unvernünftigen Umgang mit persönlichen Daten wird aktuell in der öffentlichen Diskussion darüber gestritten, wie viele Daten ein Benutzer veröffentlichen soll. Dies könnte mittelfristig dazu führen, dass für die Polyrepräsentation sinnvolle Repräsentationen nicht erfasst werden können.
Stehen dem *IR-System* ausreichend Repräsentationen in genügender Qualität zur Verfügung, stellt sich die Frage nach der Bildung der *Überschneidungen*. Wie unter anderem die Studie von Mette Skov zeigt, ist die Bildung von *Überschneidungen* zwar prinzipiell durch Maximierung der Repräsentationen implementierbar, liefert jedoch nicht notwendigerweise auch die besten Ergebnisse.

Letztendlich dürfte als Ausblick sicherlich eine größere Studie als lohnenswert erscheinen. Hierzu wäre sicherlich eine größere Internetsuchmaschine oder ein entsprechender Webkatalog ein guter Kandidat. Die Ergebnisse eines solchen praxisnahen und umfangreichen Feldversuchs könnten zur unmittelbaren Implementierung von Polyrepräsentation in bestehende Systeme führen und das Verhalten der genannten Probleme in der Praxis zeigen.

6. Literatur- und Quellenverzeichnis

Angabe der Quellen bei einer Kapitelüberschrift impliziert, daß der folgende Abschnitt auf der entsprechende Quelle basiert.

[1] Peter Ingwersen - Polyrepresentation of Information Needs and Semantic Entities – Elements of a Cognitive Theory for Information Retrieval Interaction

[2] Birger Hjørland - http://www.db.dk/bh/Core%20Concepts%20in%20LIS/articles%20a-z/ask.htm 10.09.2009

[3] National Institute of Standards and Technology - http://trec.nist.gov/ 30.09.2009

[4] Birger Larsen, Peter Ingwersen, Jaana Kekäläinen - The Polyrepresentation continuum in IR

[5] Birger Larsen, Peter Ingwersen, Berit Lund – Data Fusion According to the Principle of Polyrepresentation

[6] Mette Skov, Birger Larsen, Peter Ingwersen – Intra and Intra-Document Contexts Applied in Polyrepresentation

[7] Mentor Cana - http://www.kmentor.com/socio-tech-info/2003/06/human-information-behavior-soc.html 24.09.2009

[8] Birger Hjørland - http://www.db.dk/bh/Lifeboat_KO/CONCEPTS/polyrepresentation.htm 03.08.2009

[9] Norbert Fuhr - Information Retrieval – Ein Überblick