# Testverfahren für intelligente Indexierungs- und Retrievalsysteme anhand deutschsprachiger sozialwissenschaftlicher Fachinformation (GIRT)

## Bericht über einen Workshop am IZ Sozialwissenschaften, Bonn

## 12. September 1997, 10.30 Uhr bis 17.00 Uhr

***Gerhard Knorz***

*Der Workshop wurde gemeinsam vom Hochschulverband Informationswissenschaft, der Fachgruppe Information Retrieval der Gesellschaft für Informatik (GI-IR) und dem IZ Sozialwissenschaften veranstaltet.*

*LDV-Forum Bd. 14, Nr. 2, Jg. 1997*

*Zwischen Forschung und Praxis liegen im Bereich des Information Retrieval 20 Jahre und mehr. Wenn wir die daraus resultierende Sprachlosigkeit zwischen Forschern und Praktiker überwinden wollen, wenn wir die Anwender mit Systemen unterstützen wollen, die das gegenwärtig Gewohnte in den Schatten stellen, dann brauchen wir überzeugende Belege und erfolgreiche erste Anwendungen. Dem Workshop GIRT kann man tatsächlich zutrauen, ein erster von weiteren folgenreichen Schritten in diese Richtung zu sein.*

–Gerhard Knorz

# 1 Die Initiative „GIRT“

Seit ca. 2 Jahren wird im Hochschulverband für Informationswissenschaft (HI) und in der GI-Fachgruppe Information Retrieval über eine Initiative diskutiert, die mit dem Workshop GIRT nun eine erste Öffentlichkeit hergestellt hat. Mit dem Wechsel von Prof. Dr. Jürgen Krause von der Informationswissenschaft der Universität Regensburg an die Spitze des IZ Sozialwissenschaften in Bonn (und gleichzeitig an die Universität Koblenz an den Lehrstuhl der Software Ergonomie im Fachbereich Informatik) hat das IZ eine radikale informationstechnologische Umorientierung vollzogen, die nun auch kürzlich durch den Wissenschaftsrat ihre Bestätigung gefunden hat. Dieses Informationszentrum, das satzungsgemäß als Mitglied der GESIS (Gesellschaft Sozialwissenschaftlicher Infrastruktureinrichtungen) die sozialwissenschaftliche Forschung dokumentiert und unterstützt, hat es sich zum Ziel gesetzt, mit einer neu eingerichteten Forschungsabteilung, unterstützt durch Drittmittelprojekte und als am Markt tätiger Datenbankanbieter die technologisch-methodische Lücke zwischen informationswissenschaftlicher Forschung und karger Praxis zu schließen und sich mit eigenen Informationsangeboten an die „Spitze des Fortschritts“ (J. Krause) zu setzen. Daß gegenwärtig im Information Retrieval vieles in Bewegung gekommen ist, ist einerseits dem Internet zu verdanken und gleichzeitig einem zweifelsfrei erfolg- und einflußreichen offenen Wettbewerb für Retrievalsysteme in den USA. Dieser Wettbewerb, TREC (Text Retrieval Conference), setzt gegenwärtig die Maßstäbe für die Effektivität von Retrievalsystemen, weit über den eigentlichen Kreis der Teilnehmer aus Forschung und Industrie hinaus. NIST, der neutrale Veranstalter von TREC, stellt Dokumentsammlungen im Gigabyte-Bereich, Retrievalfragen, Wettbewerbs-

regeln und -organisation sowie Auswertungskapazität zur Verfügung – und die bisherigen Ergebnisse haben die in der Praxis dominanten Booleschen Retrievalverfahren zugunsten rankingbasierter Systeme klar aus dem Feld geschlagen.

Aus Sicht des IZ Sozialwissenschaften – wie auch sicher vieler anderer – haben die experimentellen Ergebnisse von TREC allerdings einen entscheidenden Nachteil: Erst die neuesten Ergebnisse der allerletzten Runde (TREC6, Herbst 1997) berücksichtigen (auch) die deutsche Sprache und außerdem sind – aus der Not geboren – die Dokumentsammlungen überwiegend auf Zeitungstexte konzentriert. Inwieweit also eine an englischsprachigen Zeitungen gewonnene Erkenntnis etwa auf eine deutschsprachige sozialwissenschaftliche Datenbank anwendbar ist, für die eine thesaurusbasierte intellektuelle Indexierung vorliegt, bleibt weiterhin eine Angelegenheit von Glauben und Spekulation. Schließlich zeigen gerade auch die TREC-Ergebnisse, wie sehr Fachgebiet, Topic und Textsorte in bisher nicht vorhersehbarer Weise auf die Retrievalqualität durchschlagen.

Genau hier setzt GIRT als ein Projekt des IZ Sozialwissenschaften an, das man vereinfacht als eine Art deutsches TREC bezeichnen könnte. Das IZ stellt – wie NIST – eine Testumgebung (Dokumente aus IZ-Datenbanken, Retrievalfragen, Aufbereitungs- und Auswertungskapazität) all denen zur Verfügung, die dieses Angebot zur Evaluierung ihrer Retrievalverfahren nutzen wollen. Die Vorteile haben alle Beteiligten auf *ihrer* Seite: Forschung und Entwicklung haben eine praxisrelevante Testumgebung (die bisher im deutschen Bereich fehlte), Forschung und Praxis haben vergleichbare Testergebnisse und das IZ Sozialwissenschaften hat aus erster Hand und auf eigenen Daten Belege dafür, welche Verfahren sich für seine Zwecke als überlegen herausstellen. Insbesondere kann es für seinen Fall die jahrzehntealte Streitfrage entscheiden, ob sich der Aufwand für Thesaurusentwicklung und intellektuelle Indexierung in besserer Retrievalleistung auszahlt, oder ob die manuelle Bearbeitung durch automatische Indexierungs- und Retrievalverfahren abgelöst werden kann bzw. soll.

Der nunmehr erste Workshop im September 1997 sollte die Idee und den Kontext von GIRT sowie die Ergebnisse eines Pretests mit den Systemen *freewais-sf* und *Messenger* vorstellen. Außerdem war es Ziel, GIRT mit der neuen Initiative von TREC hinsichtlich Mehrsprachigkeit unter Einschluß von Deutsch zu koordinieren.

Im gutgefüllten großen Sitzungssaal des IZ begann pünktlich die Veranstaltung.

# 2 Vorträge

## 2.1 Ziele und Perspektiven des Projektes GIRT (Krause)

In seinem kurzen Einleitungsreferat begründete und motivierte Jürgen Krause aus seiner Sicht als Initiator von GIRT das damit verbundene wissenschaftliche und praktische Interesse des IZ Sozialwissenschaften. Evaluierung von Retrievalverfahren als ein facettenreiches und in methodischer wie auch in praktischer Hinsicht höchst anspruchsvolles wissenschaftliches Problem ist ein Anliegen, das er aus Regensburg mit nach Bonn transferiert hat. Der Wechsel von der Hochschule zu einem Institut, das selbst als Informationsanbieter auftritt und den Zugang seiner Kunden zu seinen Datenbankinhalten selbst gestalten kann, stellt die Evaluierung in einen völlig neuen Kontext: Evaluierung nicht mit dem Ergebnis letztlich doch unverbindlicher Aussagen über Effektivität und Akzeptanz, sondern als konkreter Ausgangspunkt für die Gestaltung von Informationsdiensten des IZ. Diese Instrumentalisierung von Evaluierung fokussiert die Aufmerksamkeit auf Aspekte, die aus Nutzer- und Betreibersicht von besonderer Bedeutung sind: Dies sind zum einen Stellenwert und Rolle der Indexierung, die am IZ traditionell-professionell mittels eines selbstentwickelten Thesaurus intellektuell vorgenommen wird und zum andern das Interface zum Nutzer, das unbestritten die Effektivität der Recherche wesentlich bestimmt.

## 2.2 Generelle Ergebnisse der TREC-Studien, einschließlich TREC-5 (Womser-Hacker)

Frau Womser-Hacker, die sich im Rahmen ihrer im Februar dieses Jahres abgeschlossenen Habilitation intensiv mit den Ergebnissen von TREC auseinandergesetzt hat, versuchte den Stellenwert von TREC herauszuarbeiten, den Fortschritt an Entwicklung und Wissen zu charakterisieren und andererseits die Lükken aufzuzeigen, die aus dem Testdesign von TREC und der praktischen Ausgestaltung dieses Retrievalwettbewerbs folgen. Besonders interessant und glücklich für die Diskussion war es, daß mit Peter Schäuble von der ETH Zürich, einem der seit Anbeginn bei TREC beteiligten Teilnehmer (und Koordinator des sich

nun entwickelnden europäischen TREC-Ablegers), ein Insider zugegen war. Auf diese Weise gewann das Thema über die Systematik hinaus ein unmittelbares und authentisches Element, das einen Workshop immer aufwertet.

Zunächst stellte Womser-Hacker Evaluierung als ein Kernthema des Information Retrieval heraus, dessen Geschichte sie in 3 Phasen mit jeweils eigenem Erkenntnisgewinn und Entwicklungsstand einteilte:

1.) **1955–1980** mit den Cranfied-Tests und den Experimenten mit SMART und Medlars: In dieser Phase etablierte sich das Bewußtsein, daß Standardisierung und Wissenschaft, aber auch viel harte Arbeit für die Bewertung von Retrievalverfahren notwendig sind. Man lernte, das Mögliche im Bereich von 40%–60%, was Precision und Recall als die Standardmaße betrifft, einzuschätzen, und formulierte deren inverse Beziehung als empirisches Gesetz. Als offene Frage verblieb, wie sehr die Ergebnisse vom Kontext abhängig sind und warum die Systeme gerade so gut (oder schlecht) sind, wie sie sind. Und bereits in den Anfängen öffnete sich die so viel beklagte Kluft zwischen Praxis und Forschung.

2.) **1980–1990** als eine Phase mit speziellem Bezug zur deutschen Szene mit Projekten wie PADOK (Regensburg), AIR (Darmstadt) und LIVE (Berlin). Standardisierung im Testdesign und Operationalisierung von Systemeigenschaften kennzeichnen den Fokus. Das Bewußtsein für statistisch und meßtheoretisch „saubere“ Bedingungen und Ergebnisse hat sich entwikkelt, ebenso wie auch für die Problematik der Konzepte von Relevanz, Aboutness und Semantik. Methoden zur Schätzung schwer meßbarer Parameter wurden entwickelt, die Mehrfachnutzung von Testkollektionen gewann an Bedeutung, die Testtypen Experiment und Untersuchung differenzierten sich aus und es zeigte sich, daß mit zunehmender Komplexität der Systeme die Evaluierung immer schwieriger wird.

3.) **ab 1990** dominieren TREC und seine Ausdifferenzierungen (special-tracks). GIRT soll sich hier mit einordnen.

TREC (Text Retrieval Conference) hat seit dem November 1992, dem Startpunkt, mit TREC 1 eine außerordentliche Ausstrahlung erreicht und ist zweifellos für das Gebiet des Information Retrieval ein großer Gewinn. Der aktuelle Stand veröffentlichter Ergebnisse ist gegenwärtig TREC 5. TREC 6 (mit dem Schwerpunkt Crosslingual Retrieval) hat 1997 bereits stattgefunden. Interessierte können alle Veröffentlichungen einsehen unter dem URL http://www-nlpir.nist.bov/trec/ . Plakative Basisinformation über TREC, speziell auch über das Testdesign, finden sich als Folienfolge unter http://www.iud.fh-darmstadt.de/iud/wwwmeth/publ/slide/owfrtr1.htm.

TREC adressiert als Zielgruppe den „dedicated searcher", unterscheidet grundsätzlich zwischen adhoc-Abfragen und Routing sowie zwischen automatischer, manueller und interaktiver Entwicklung der Suchanfrage. Grundsätzlich wird von einem „gerankten" Ergebnis ausgegangen. Das Evaluierungsprogramm (und damit das Erkenntnisinteresse) entwickelt sich über die Zeit, so daß sich die Testbedingungen der einzelnen Runden unterscheiden und die Ergebnisse nur sehr eingeschränkt direkt vergleichen lassen. Spezifische Probleme (Filtering, fehlerhafter Input, Interaktivität, …) werden in special tracks behandelt. Die Testkollektionen sind heterogen, konstruiert und „groß" (Gigabyte-Bereich), vornehmlich bestehend aus Zeitungstexten. Die Fragekollektionen umfassen jeweils 50 Suchproblemstellungen. Die Teilnehmeranzahl wächst (TREC1 25 Systeme, TREC4 35 Systeme) und umfaßt Forschungsprototypen wie auch Systeme von kommerziellen Herstellern.

Die Ergebnisse der einzelnen TREC-Runden lassen sich recht gut in eine Entwicklungslinie einpassen:

1.) **TREC 1:** Fast alle Teilnehmer hatten mit Effizienzproblemen zu kämpfen, aber es zeigte sich, daß system- und organisationsseitig das TREC-Konzept praktikabel war.

2.) **TREC 2:** Die Ergebnisse waren deutlich verbessert. Ausschlaggebender Erfolgsfaktor war die Optimierung der Termgewichtungen.

3.) **TREC 3:** Systeme und Experimente wurden komplexer. Hybride Ansätze führten zu Ergebnisverbesserungen.

4.) **TREC 4:** Erfolgsfaktor war nunmehr die Einbeziehung linguistischer Analysen (Behandlung von Phrases, Frageerweiterung).

5.) **TREC 5:** Verbesserung der Resultate wurden erreicht, indem man das Problem der Längennormalisierung in den Griff bekam (Berücksichtigung der Tatsache, daß das Auftreten eines Suchwortes in einer 10-Zeilen-Nachricht einen anderen Stellenwert hat als in einem 10-Seiten-Aufsatz.)

Der Shift in den Erfolgsfaktoren ist z. T. auf einen Wechsel der Testbedingungen und damit der Anforderungen zurückzuführen: Insbesondere wurden die anfangs sehr ausdifferenzierten Frageformulierungen radikal gekürzt. Dadurch wurde die Aufgabe deutlich schwieriger mit dem Effekt, daß die Retrievaleffektivität signifikant zurückging und nunmehr das Problem der Frageerweiterung in den Vordergrund rückte.

Ein interessantes und bedenkenswertes Ergebnis im Hinblick auf die Frage nach der Übertragbarkeit der Resultate lieferte ein Signifikanztest bei den TREC 3-Ergebnissen: Die Varianz durch die verschiedenen Such-Topics erwies sich als höher als die Varianz verschiedener Retrievalsysteme.

Insgesamt liefert TREC bisher folgende Aussagen:

- TREC hat die Entwicklung von Retrievalverfahren enorm befruchtet. Die heutigen Verfahren, in Konkurrenz zu den früheren auf den „alten“ Testdaten, liefern deutlich verbesserte Ergebnisse.
- Viele Systeme erreichen denselben Leistungsstandard. Unterschiedliche statistische Modelle schlagen wenig auf die Effektivität durch, genauso wie elaborierte Modelle gegenüber einfachen Ansätzen keine signifikante Verbesserung nachweisen.
- Stark linguistisch orientierte Verarbeitung rechtfertigt den Mehraufwand nicht.
- Elaborierte manuelle Frageentwicklung ist keineswegs immer besser als einfache (automatische).
- Anwendung von Relevance Feedback-Techniken zeigt eindeutige Verbesserungen.
- Data Fusion ist eindeutig positiv.

- Der Aufwand in eine verbesserte Behandlung der Anfrage schlägt besser auf Effektivität durch als eine bessere Dokumentenanalyse.
- Mit TREC 5 änderte sich die vorher bestehende Situation, daß sich die Ergebnisdokumentmengen der einzelnen Systeme (sowohl was die relevanten, als auch was die nicht-relevanten Dokumente angeht) erstaunlich wenig überschneiden.

## 2.3 Ergebnisse des GIRT-Pretests (Kluck)

Michael Kluck, am IZ Sozialwissenschaften für die Durchführung von GIRT zuständig, erläuterte Test-Design und Ergebnisse des ersten Tests. Ziel war es im wesentlichen gewesen, die Praktikabilität des Testkonzeptes zu erproben und Erfahrungen mit Organisation und Ablauf zu sammeln. Schließlich lassen sich im Vorfeld die für den praktischen Aufwand wichtigen Fragen nach der durchnittlichen Dauer einer Recherche von Testpersonen, nach deren praktischen Problemen oder nach der zu erwartenden Anzahl von Trefferdokumenten nur unter großer Unsicherheit prognostizieren.

Der Pretest wurde von zwei Systemen bestritten: *freewais-sf*, eine einfache, verbreitete Suchengine im Internet mit statistischem Ranking und einfacher regelbasierter Morphologie sowie *Messenger*, das (konventionelle) Standardsystem für eine Suche nach manuell indexierten IZ-Dokumenten, in einer kategorisierten bibliographischen Datenbank mit Freitextsuche

Die Testkollektion umfaßt ca. 15.000 Dokumente der beiden sozialwissenschaftlichen Datenbanken SOLIS und FORS mit Titel und Abstract. Für eine kleine Untermenge können auch Volltexte zur Verfügung gestellt werden. 9 Anfrageprobleme waren als Standardtestfragen vorgegeben und die zugehörige Relevanzbeurteilung war durch einen IZ-Juror bereits im voraus durch exhaustive Recherchen vorgenommen worden. Grundsätzlich sollen auch Spezialfragen möglich sein, die von den Testteilnehmern eingebracht werden. Voraussetzung ist, daß sie dokumentiert und veröffenlicht sind. Die Relevanzbeurteilung findet dann im Rahmen der Testauswertung statt.

Die Testpersonen waren 8 informationswissenschaftlich vorgebildete Personen mit und ohne Erfahrung im Online-Retrieval. Jede Person bearbeitete alle 9 Fragen mit beiden Systemen (mit wechselnder Reihenfolge), wobei ein Zeitbudget von ca. 2–4 Stunden benötigt wurde. Um technische Probleme und überhaupt die Interface-Problematik aus dem Test auszuklammern, stand als technische Vermittlung ein professioneller Rechercheur als Bediener der Tastatur zur Verfügung.

Eine Festlegung, die sich als problematisch herausgestellt hat, war die Begrenzung der Antwortmengen auf 30 Treffer. Diese Restriktion sollte den Aufwand kalkulierbar machen, hat aber in unzulässiger Weise das Suchverhalten und die Ergebnisse beeinflußt.

Die Relevanz von Dokumenten wurde auf einer 4-stufigen Skala bewertet und für Precision-Recall-Auswertungen auf eine binäre Relevanzentscheidung abgebildet. Die Konsistenz von Urteilen unterschiedlicher Juroren lag bei 70–80%.

Die Ergebnisse des Pretests liegen detailliert als Precision/Recall-Diagramme vor und sollen aufgrund der geringen Anzahl von Fragen sowie der unzweckmäßigen Begrenzung der Antwortmengen nicht überinterpretiert werden. Dennoch ergeben sich eine Reihe interessanter Beobachtungen:

- Die Überschneidungen der Antwortmengen liegen insgesamt auf sehr niedrigem Niveau: nur 21% der relevanten Dokumente sind in mehr als einer einzigen Recherche gefunden worden, und das bei jeweils 8 * 2 = 16 Recherchen/Frage!
- Es gibt keinen klaren Anhaltspunkt dafür, daß eines der Systeme dem anderen überlegen ist. Sollte sich dieses anhand von 9 Fragen und 72 Recherchen gewonnene Ergebnis bestätigen, so bedeutet dies, daß die manuelle Indexierung vom Kunden nicht mit Gewinn genutzt wird. Der Nutzer hat allerdings eine klare Meinung darüber, mit welchem System er besser zurechtgekommen ist und die besseren Ergebnisse erzielt hat: nämlich mit dem Booleschen System. Dieser subjektive Eindruck widerspricht der objektiven Beobachtung.
- Eine detaillierte Analyse der verwendeten Frageformulierungen ergibt, daß vielfach zentrale Aspekte der vorgegebenen Suchprobleme unberücksichtigt geblieben sind. Es drängt sich der Eindruck auf, daß ein simples Abschreiben der vorgegebenen Suchformulierung bei *freewais-sf* zu besseren Ergebnissen hätte

führen können als das Ergebnis eigenen Nachdenkens. Daß niemand überhaupt auf diese Idee gekommen ist, zeigt, daß die Testpersonen mit dem Konzept statistisch basierten Retrievals nicht vertraut waren.

- Den Testpersonen fehlte auch weitgehend das Bewußtsein, daß Boolesches Retrieval unter Verwendung von Deskriptoren die Orientierung im Vokabular nahelegt. Nur wenige Testpersonen mit Retrievalerfahrung, aber kein Laie (!) haben den bereitliegenden Thesaurus zur Suche benutzt.
- Rechercheure der IZ Sozialwissenschaften erzielen deutlich bessere Rechercheergebnisse als die Testpersonen. Hier schneidet *Messenge*r als das gewohnte und die Vorgehensweise prägende Recherchesystem klar besser ab als *freewais-sf*.

## 2.4 Multilingualität in TREC (Schäuble)

Hintergrund für den Vortrag von Peter Schäuble von der ETH Zürich war die Neuerung im Rahmen von TREC 6 (1997), CrossLanguage Information Retrieval (CLIR) mit den Sprachen Deutsch, Englisch, Französisch und prinzipiell auch Italienisch in die Evaluierung aufzunehmen. Zur Vorbereitung dieses neuen Schwerpunktes war eine mit Europäern und US-Forschern besetzte Arbeitsgruppe ins Leben gerufen worden, deren europäischer Sprecher Peter Schäuble ist. In Weiterentwicklung dieser Aktivitäten ist für 1998 ein European TREC mit Unterstützung von CEPIS, GI, BCS und NIST geplant. Und genau dies war der konkrete Ansatzpunkt für die nachfolgende Schlußdiskussion und für das Interesse, konkrete gemeinsame Perspektiven zu entwickeln.

Wer nun bei eher strategisch/politischem Hintergrund einen tendenziell deskriptiven, farblosen Vortrag erwartet hätte, der hatte sich absolut verrechnet. Schäuble vermittelte in konzentrierter Form einen ausgezeichneten Überblick über die gegenwärtige Landschaft der Ansätze für ein sprachgrenzenüberschreitendes Information Retrieval. Die grundsätzliche Idee der einzelnen erfolgreichen Konzepte wurde gut nachvollziehbar dargestellt. Die Teilnehmer des Workshops

waren – dem an lebendiger Diskussion gezeigten Interesse zufolge – beeindruckt von der – man kann es so sagen – Raffinesse des an der ETH Zürich verfolgten Ansatzes. Und vor allen Dingen von den vorgestellten Ergebnissen!

Zunächst einmal sind die möglichen Anwendungen von Crosslingual IR eines kurzen Nachdenkens wert: Für mehrsprachige Länder (für jemanden aus Zürich ein Heimspiel!) und Organisationen liegt der Bedarf auf der Hand. Darüber hinaus bietet das Web mit seiner englischen Standardsprache das Beispiel für einen weiten Nutzerkreis mit großem passivem aber eher kleinem aktiven Wortschatz in einer Fremdsprache. In professionellem Kontext kann man aber auch an monosprachliche Nutzer denken, für die eine maschinelle Rohübersetzung eines vorher gefundenen fremdsprachlichen Dokumentes zumindest eine Relevanzentscheidung ermöglicht. Eine besonders interessante Anwendung stellt die Recherche von Bildern auf der Basis von Bildbeschreibungen dar, deren Originalsprache den Suchenden gar nicht interessieren muß.

Nachdem die Probleme „naiver Ansätze" diese als untauglich disqualifiziert hatten, gab Schäuble einen kurzen Abriss der Geschichte einschlägiger wissenschaftlicher Arbeiten, beginnen bei Salton 1970 (wie könnte es anders sein?) und eine systematische Klassifikation prinzipiell möglicher Retrievalansätze, die Sprachgrenze zu überschreiten. (Eine entsprechende graphische Darstellung findet sich unter http://www.iud.fh-darmstadt.de/iud/wwwmeth/publ/slide/clirva1.htm.)

Auf zwei sehr unterschiedliche Ansätze ging Schäuble im Detail ein: zum einen auf „Latent Semantic Indexing" (LSI) und auf den an der ETH Zürich verfolgten Ansatz, der einen Ähnlichkeitsthesaurus berechnet und für das Retrieval anwendet. Hinter LSI steckt die Idee, daß man Dokumente und Terms jeweils auf einen gemeinsamen abstrakten Vektorraum abbilden kann, der sehr viel weniger Dimensionen hat als etwa der ursprüngliche Dokumentenraum des Vektormodells. Dieser reduzierte abstrakte Raum ist, natürlich, sprachunabhängig.

Ausführlicher will ich den Ansatz unter Einbeziehung eines Ähnlichkeitsthesaurus nachzeichnen: Schäuble lag viel daran, herauszustellen, daß Dokumentenräume und Termräume sich dual verhalten. In einem Dokumentenraum beschreibt man Dokumente über die Terms (das Vokabular), die sie enthalten und in einem Termraum werden Terms über die Dokumente spezifiziert, in denen sie vorkommen. Jede wahre Aussage über Zusammenhänge im Dokumentenraum läßt sich mechanisch in eine entsprechende wahre Aussage im korrespondierendem Termraum transformieren. Mit dieser interessanten Sichtweise läßt sich neu begründen, was im Information Retrieval eine lange und keineswegs flächendek-

kend erfolgreiche Geschichte hinter sich hat: die Assoziationsfaktoren oder anders benannt: der Ähnlichkeitsthesaurus. Terms werden demnach genau dann zueinander in Beziehung gesetzt, wenn sie häufig gemeinsam in Dokumenten auftreten.

Ein Ähnlichkeitsthesaurus kann zunächst zur Frageerweiterung eingesetzt werden: Die Suchterms werden um weitere Terms aus dem Ähnlichkeitsthesaurus angereichert. Zu einem CLIR-Ansatz wird dieses Verfahren dann, wenn man eine zweisprachige Dokumentenkollektion mit korrespondierenden Dokumenten zur Verfügung hat: Man berechnet dann die Ähnlichkeit eines Terms zu den Terms in jeweils anderssprachigen Dokumenten. So kann man etwa eine deutsche Anfrage über den Ähnlichkeitsthesaurus in eine Menge gewichteter französischer Suchterms abbilden, mit denen man dann in französischsprachigen Dokumenten recherchiert. Wichtig zu wissen ist, daß die französischen Suchterms keinesfalls Übersetzungen sein müssen bzw. sein sollten. Vielmehr geht es (nur) darum, solche französische Wörter zu finden, wie sie in Dokumenten auftreten, in deren deutschen Übersetzung das deutsche Suchwort vorkommt. Damit ist das Problem deutlich einfacher, als es z. B. in einem Übersetzungssystem gelöst werden muß.

Nun erscheinen Dokumentenkollektionen, in denen alle Dokumente in zwei Sprachen verfaßt sind, sehr selten. Die Züricher sind dennoch fündig geworden: eine erste Anwendung lieferte eine Dokumentensammlung zum Schweizer „systematischem Recht“ in Deutsch und Französisch. Anhand von 53.000 Dokumenten wurde ein Ähnlichkeitsthesaurus entwickelt, der anschließend u. a. erfolgreich für die deutschsprachige Recherche in französischsprachigen Gerichtsurteilen (Wechsel der Textsorte!) angewandt wurde. Um die detaillierten Evaluierungsergebnisse zusammenzufassen: Gegenüber einer monosprachigen einfachen Recherche *ohne* Ähnlichkeitsthesaurus bringt die Crosslingual-Recherche *mit* Ähnlichkeitsthesaurus eine deutliche Qualitätssteigerung, und letztlich kann man sagen, daß der Sprachwechsel dem Verfahren nur einen Effektivitätsverlust unter 5% kostet

Der besondere Clou ist den Zürichern allerdings damit gelungen, daß sie das Verfahren auch ohne vorliegende zweisprachige Dokumentenkollektion implementieren konnten. Sie stellten eine solche Sammlung „einfach“ selbst her: Sie filterten Agenturmeldungen eines sehr langen Zeitraums (mehr als 10 Jahre) und glichen verschiedensprachige Meldungen anhand einfacher Kriterien wie Datum, Klassifikation, gemeinsames Auftreten sprachunabhängiger Zeichenfolgen (Namen) ab. Daß die so ermittelten Dokumentenpaare nicht notwendigerweise

direkte Übersetzungen voneinander sind, muß aus praktischer und theoretischer Sicht nicht stören. Den damit berechneten Ähnlichkeitsthesaurus setzte die ETH Zürich dann in den Experimenten für TREC 6 ein. Wie sehr man umdenken muß, wenn man versucht, das Potential eines solchen Verfahrens einzuschätzen, machte Schäuble an einem Exempel plastisch deutlich: Nimmt man etwa den Schweizer Parlamentsthesaurus in die Hand, so hat man es mit 2 cm bedruckten Seiten zu tun. In gleicher Form ausgedruckt ergibt der bei TREC6 eingesetzte Ähnlichkeitsthesaurus einen 70 km hohen Papierstapel!

# 3 Abschlußdiskussion und Resümee

Die Abschlußdiskussion, von Gerhard Knorz (FH Darmstadt) moderiert, sollte im wesentlichen 2 Fragen klären:

1. Gibt es Interesse an und Anforderungen für das Angebot, das GIRT all denjenigen macht, die ein deutschsprachiges Retrieval anbieten und die dieses Retrieval evaluieren wollen?
2. Lassen sich die Initiativen für GIRT und für ein europäisches TREC in einen gemeinsamen Rahmen einbetten?

Der erste Punkt wurde insofern nur kurz behandelt, als übereinstimmend festgestellt wurde, daß eine Evaluierungsumgebung, wie sie GIRT anbietet, ein Desiderat für jeden darstellt, der Retrievalverfahren für deutschsprachige Texte entwickelt bzw. optimiert. Gegenüber den Planungen weitergehende Anforderungen wurden nicht formuliert.

Der zweite Punkt war im Verlauf des Workshops bereits häufiger andiskutiert worden. Im wesentlichen wurde zwischen Krause als dem Ausrichter von GIRT und Schäuble als dem europäischen Sprecher der Arbeitsgruppe für ein europäisches TREC geklärt, daß das IZ Sozialwissenschaften unbestritten und willkommen die neutrale Rolle für ein europäisches TREC spielen kann, wie sie NIST in den USA für TREC innehat. Daß die Verwaltung weiterer Dokumenten- und Fragekollektionen mit dem damit verbundenen Auswertungsaufwand nicht mit den

Mitteln des IZ getragen werden kann, wird klar akzeptiert und sollte kein Problem darstellen, da eine Projektförderung für ein solches Vorhaben sicher erreichbar scheint.

So hatte der Workshop allen Teilnehmern eine Menge zu bieten gehabt: Interessante Informationen, lebhafte Diskussionen und ein sehr vielversprechendes praktisches Ergebnis. Der Aufschwung des Gebietes „Information Retrieval“ setzt sich fort!