Evaluierung von Anfragetransfers für sozialwissenschaftliche Internetdokumente

Robert *Strötgen*, Hildesheim und Udo *Riege*, Bonn

1 Kontext: Behandlung semantischer Heterogenitä

In digitalen Bibliotheken als integrierten Zugängen zu in der Regel mehreren verschiedenen Dokumentsammlungen tritt Heterogenität in vielerlei Spielarten auf:

- als technische Heterogenität durch das Zusammenspiel verschiedener Betriebs-, Datenbank- oder Softwaresysteme,
- als strukturelle Heterogenität durch das Auftreten verschiedener Dokumentstrukturen und Metadaten-Standards und schließlich
- als semantische Heterogenität, wenn Dokumente mit Hilfe unterschiedlicher Ontologien (hier verwendet im weiteren Sinn von Dokumentationssprachen wie Thesauri und Klassifikationen) erschlossen wurden oder aber Dokumente überhaupt nicht mit Metadaten ausgezeichnet wurden.

Semantische Heterogenität lässt sich behandeln, indem die Standardisierung von Metadaten (z.B. von der Dublin Core Metadata Initiative[1] oder das *Resource Description Framework*[2] (RDF) im Kontext des *Semantic Web*[3]*)* vorangetrieben und ihre Verwendung gefördert wird. Allerdings besteht auf Grund der unterschiedlichen Interessen aller beteiligten Partner (u.a. Bibliotheken, Dokumentationsstellen, Datenbankproduzenten, „freie“ Anbieter von Dokumentsammlungen und Datenbanken) kaum die Aussicht, dass sich durch diese Standardisierung semantische Heterogenität restlos beseitigen lässt. (Krause 2003) Insbesondere ist eine einheitliche Verwendung von Vokabularen und Ontologien nicht in Sicht.

Im Projekt CARMEN[4] wurde unter anderem das Problem der semantischen Heterogenität einerseits durch die automatische Extraktion von Metadaten aus Internetdokumenten (Strötgen & Kokkelink 2001) und andererseits durch Systeme zur Transformation von Anfragen über Cross-Konkordanzen und statistisch erzeugte Relationen angegangen. (Hellweg et al. 2001) Ein Teil der Ergebnisse der Arbeiten am IZ Sozialwissenschaften[5] waren statistische Relationen zwischen Deskriptoren, die mittels Kookurrenzbeziehungen berechnet wurden. Diese Relationen wurden dann für die Übersetzung von Anfragen genutzt, um zwischen verschiedenen Ontologien oder auch Freitexttermen zu vermitteln (siehe Abbildung 1). Das Ziel dieser Übersetzung ist die Verbesserung des (automatischen) Überstiegs zwischen unterschiedlich erschlossenen Dokumentbeständen, z.B. Fachdatenbanken und Internetdokumenten, als Lösungsansatz zur Behandlung semantischer Heterogenität.

1. http:// dublincore.org/
2. http://www.w3.org/RDF/
3. ttp://www.w3.org/2001/sw/
4. Sonderfördermaßnahme im Rahmen von Global-Info (*C*ontent *A*nalysis, *R*etrieval and *M*etaData: *E*ffective *N*etworking), http://www.mathematik.uni-osnabrueck.de/projects/carmen/
5. http://www.gesis.org/iz/

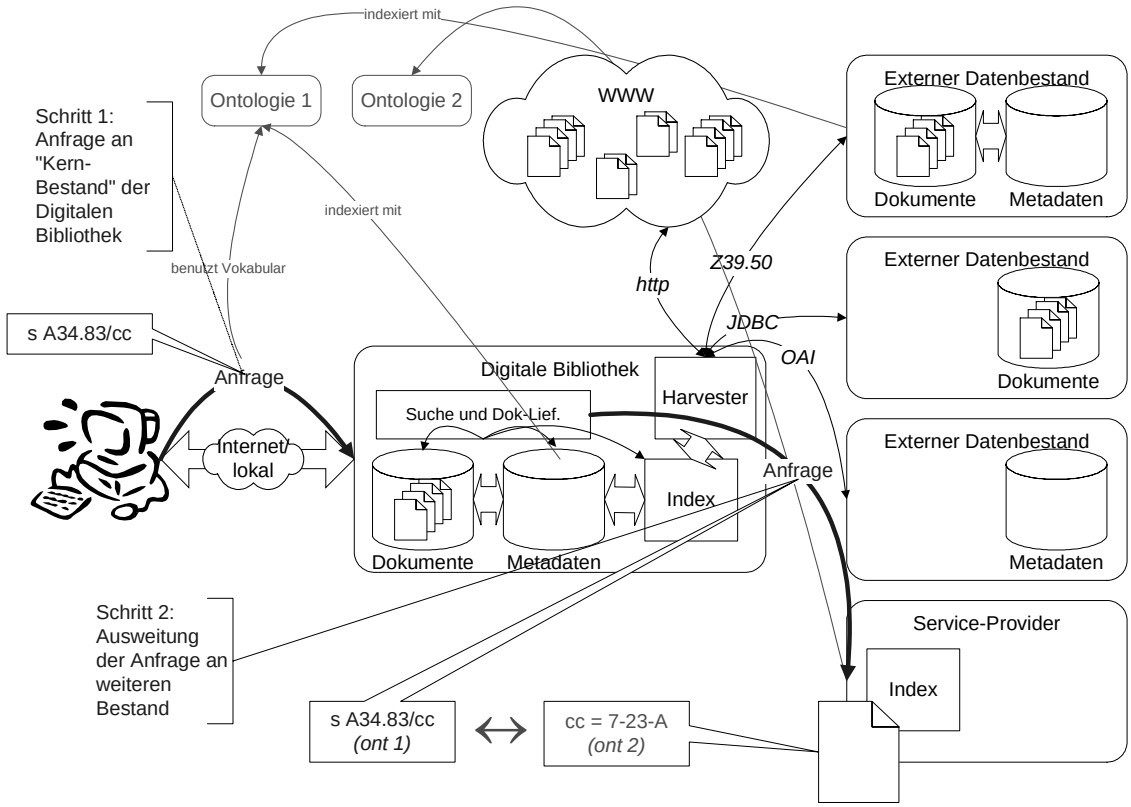

Abbildung 1: Beispiel-Szenario für semantische Heterogenität in digitalen Bibliotheken

In der Evaluierung der Verfahren zeigte sich, dass die Übersetzung von Anfragen mit Hilfe statistischer Relationen prinzipiell zu einer Verbesserung der Retrievalqualität führen konnte. Insbesondere wurde eine Verbesserung des Recall erreicht, der Ballast wurde dabei mehr oder weniger stark vergrößert. Eine systematische Analyse der Bedingungen für erfolgreiche Anwendungen konnte jedoch noch nicht geleistet werden. (Strötgen 2002)

Es wurden allerdings Ansatzpunkte aufgezeigt, wie das Verfahren noch verbessert werden könnte. An dieser Stelle setzt das Projekt ASEMOS[6] an, das die in CARMEN entwickelten Ansätze weiter verfolgt. Verbesserungen wurden besonders durch ausgefeilte Verfahren des maschinellen Lernens zur Simulation von Doppelkorpora und durch die Anwendung linguistischer Verfahren für die Verwendung „echter“ Freitextterme angestrebt. (Strötgen 2004)

Die Relevanz der Problematik hat sich auch nach Abschluss der Arbeiten am Projekt CAR-*MEN* zunehmend verdeutlicht. Integrierte Portale im Kontext digitaler Bibliotheken wie z.B. die Wissenschaftsportale Infoconnex[7] und Vascoda[8] planen den Einsatz entsprechender Anfrageübersetzungstechniken oder haben sie bereits eingeführt. (Stempfhuber 2004:243) Auch im Kontext des Semantic Web wird die Vermittlung zwischen verschiedenen Ontologien als ein zentrales Problem der Integration heterogener Dokumente angesehen und bearbeitet. Im näheren Kontext dieser Arbeit wird an der Universität Hildesheim bei dem

6. *A*pplied *S*tatistics, *E*valuation and *M*achine Learning: *O*ntology *S*witching
7. http://www.infoconnex.de/
8. http://www.vascoda.de/

virtuellen Bibliotheksregal MyShelf eine Vermittlung zwischen Klassifikationen durch Ontology Switching angewendet. (Mandl & Womser-Hacker 2002, Kölle et al. 2004)

2 Evaluierung

Im Projekt *CARMEN* konnte nur eine auf wenige Beispielanfragen beschränkte Evaluierung durchgeführt werden. (Binder et al. 2002) Für die systematischere Weiterentwicklung des Verfahrens zur Erzeugung semantischer Relationen und für eine genauere Bewertung der Auswirkungen auf das Retrievalergebnis ist es im Projekt *ASEMOS* erforderlich, die Evaluierung mit höherem Aufwand zu betreiben. Dies betrifft sowohl Testkorpora und Testanfragen als auch die Erstellung von Relevanzurteilen und die Auswertung der Ergebnisse.

2.1 Testkorpora

In *CARMEN* wurde im Jahr 2001 durch intellektuelle Auswahl und manuelles Laden ein Testkorpus erstellt, das 7915 HTML-Dokumente (neben einer Reihe von Dokumenten in anderen Formaten wie PDF, PostScript, Microsoft Word) mit sozialwissenschaftlich relevanten Inhalten zu den drei Themengebieten Frauenforschung, Migration und Betriebssoziologie enthält. Die gesammelten Dokumente sind überwiegend in deutscher Sprache verfasst und umfassen formal verschiedenste Texttypen wie Projektbeschreibungen, Veranstaltungsankündigungen, Aufsätze oder Mitarbeiterlisten. Alle Dokumente sind im Internet frei verfügbar und in aller Regel mit keinerlei Metadaten versehen.[9]

Dieses Korpus war nicht nur als „simuliertes Parallelkorpus“ (Binder et al. 2002:37) Grundlage für die Berechnung semantischer Relationen, sondern lieferte auch die Dokumente für die spätere Evaluierung. Systematische Probleme bei der Bewertung der Ergebnisse sind daher nicht auszuschließen. Aus diesem Grund wurde im Projekt *ASEMOS* für die weitere Evaluierung ein zweites Korpus erstellt, das ausschließlich für die Evaluierung genutzt wird.

Dabei wurde nicht erneut ein intellektuelles Verfahren zur Auswahl von Internetquellen gewählt, sondern auf der Grundlage des Open-Source-Spiders *Jobo*[10] ein speziell angepasster Webroboter implementiert. Dieser wurde auf elf Startseiten (vor allem einschlägige Linklisten) zu den gleichen Themenbereichen wie in *CARMEN* angesetzt und sammelte unter Berücksichtigung einiger besonderer Heuristiken (Dokumente aus der Linkliste werden gespidert, aber nicht in der Datenbank abgespeichert, um die Datenbank nicht mit kommentierten Beschreibungen von Internetdokumenten zu füllen) innerhalb einer knappen Woche 7095 überwiegend deutschsprachige HTML-Dokumente ein. Diese wurden dabei mit dem Werkzeug Tidy[11] syntaktisch bereinigt und in eine PostgreSQL[12]-Datenbank abgespeichert.

2.2 Testanfragen

In *CARMEN* wurden für jedes der drei Themengebiete lediglich zwei Testanfragen evaluiert. (Binder et al. 2002:42ff) Ganz offensichtlich lassen sich dadurch nur exemplarische Ergebnisse beschreiben. Immerhin ließen sich damit Beispiele für die Auswirkungen der Anfrage-

9. Eine genauere Beschreibung des Korpus findet sich in Binder et al. 2002:5-6.
10. http://www.matuschek.net/software/jobo/
11. http://tidy.sourceforge.net/
12. http://www.postgresql.org/

transfers beschreiben und Tendenzen erkennen. Für eine genauere Bewertung sind diese insgesamt sechs Anfragen aber zu wenig.

Für die weitere Bewertung wurde im Projekt *ASEMOS* die Zahl mit sieben neuen Anfragen auf insgesamt dreizehn Anfragen mehr als verdoppelt. Auch wenn eine deutlich höhere Anzahl an Testanfragen wünschenswert wäre, ist der damit verbundene Aufwand bei der Erarbeitung von Relevanzurteilen (dazu im Folgenden mehr) in diesem Kontext nicht zu leisten.

Die Test-Anfragen wurden sowohl für *CARMEN* als auch für *ASEMOS* mit einem statistischen Suchwerkzeug ausgeführt. In *CARMEN* wurde dafür das kommerzielle Produkt *FULCRUM* verwendet, in *ASEMOS* die freie Open-Source-Suchmaschine *Lucene*.[13] Um eine Vergleichbarkeit mit den Ergebnissen aus *CARMEN* zu gewährleisten, wurden alle alten Anfragen aus CARMEN mit Lucene wiederholt.

2.3 Relevanzurteile

Für die Beurteilung der Retrievalqualität im Sinn einer Effektivitätsmessung auf der Grundlage von Recall und Precision sind Relevanzurteile erforderlich. (Womser-Hacker 1989:27ff) Dabei wurde eine binäre Unterscheidung zwischen „relevant“ und „nicht relevant“ gewählt. Die Beurteilung der Relevanz wurde dabei von verschiedenen Mitarbeiterinnen und Mitarbeitern an der Universität Hildesheim und am IZ Sozialwissenschaften intellektuell geleistet.[14]

Diesen Juroren wurde ein webbasiertes Werkzeug zur Relevanzbewertung bereitgestellt. Dieses JSP- und Servlet-basierte Werkzeug integriert sich nahtlos als ein Modul in das in Java implementierte *ASEMOS*-Gesamtsystem. In diesem Informationssystem sind linguistische und semantische Analyse, maschinelles Lernen und automatisches Klassifizieren, Berechnung semantischer Relationen, Indexierung von Dokumenten, Transfer und Ausführung von Anfragen integriert, alle Module benutzen ein gemeinsames Datenmodell. Diese Vorteile rechtfertigen den Implementierungsaufwand gegenüber der Nutzung eines existierenden Systems für die Relevanzbeurteilung bei Retrievaltests.

Zur Relevanzbewertung wird den Juroren eine Liste aller noch unbewerteten Dokumente zu einer Anfrage präsentiert. Die Reihenfolge der Dokumente wurde dabei über eine Zufallsfunktion gemischt, um zu verhindern, dass die Juroren am Ende der Liste nicht relevante Dokumente erwarten und dies ihre Relevanzentscheidung beeinflusst. Die Dokumente werden jeweils aus der Datenbank ausgelesen und in einem neuen Browserfenster geöffnet.

In der Datenbank werden alle Relevanzurteile den jeweiligen Themen zugeordnet, nicht der Anfrage. Da über den Anfragetransfer zu einem Thema eine Vielzahl von Anfragen erzeugt wird, können über die Pooling-Methode Relevanzurteile zu einem Thema vielfach verschiedenen zugehörigen Anfragen zugeordnet werden.

Für den Bereich Sozialwissenschaften wurden bislang zu insgesamt 1004 Anfragevarianten der dreizehn Ausgangsfragen 16175 einzelne Relevanzurteile für die Dokumente in den

13. http://lucene.apache.org/
14. An dieser Stelle sei vor allem Frau Susanne Rauch vom IZ Sozialwissenschaften und Frau Kathrin Wünnemann von der Universität Hildesheim für die unermüdliche Arbeit als Jurorinnen gedankt!

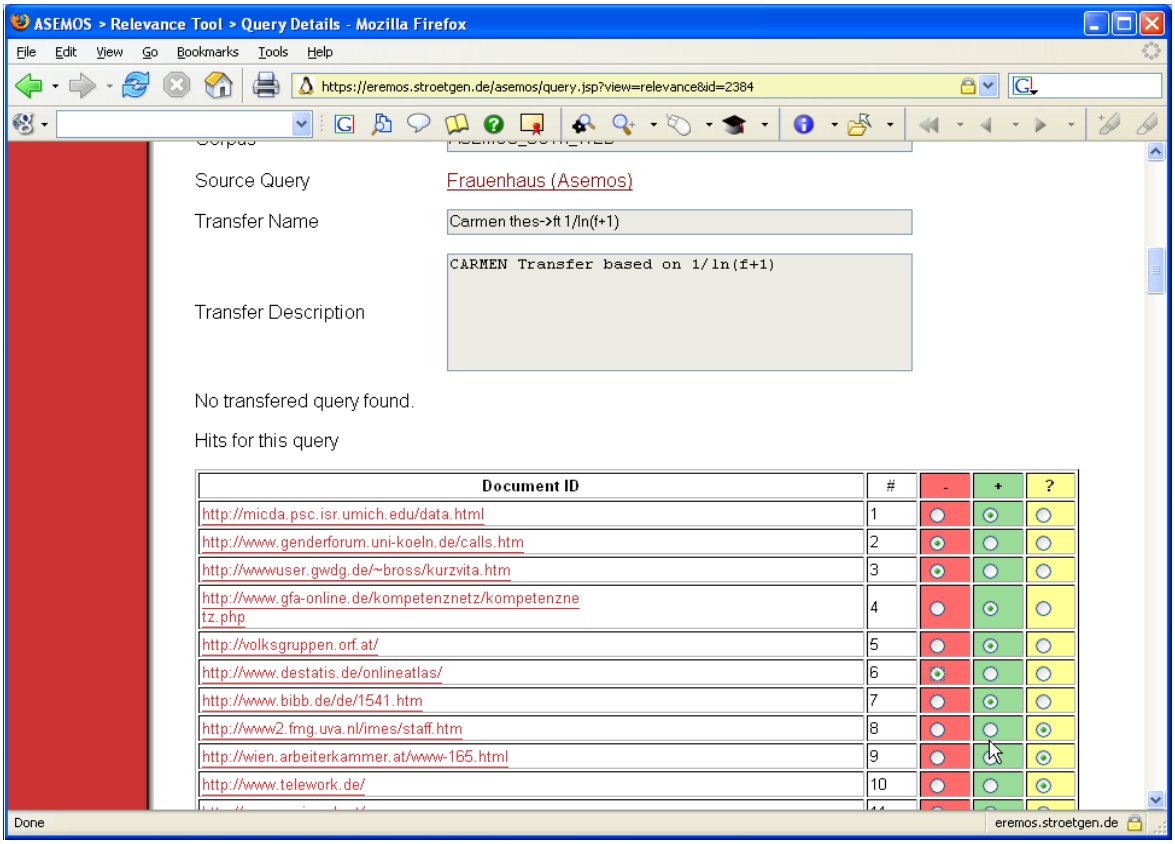

Abbildung 2: Relevanzbewertung mit dem ASEMOS Relevance Tool

Testkorpora erarbeitet. Selbst wenn die geübten Juroren je nach Dokument teilweise mehrere Relevanzurteile in einer Minute treffen können, ist der Aufwand immens.

2.4 Auswertung

Bei der Beurteilung der Relevanz wird in *ASEMOS* nicht mehr nur die gesamte Treffermenge, sondern auch das Ranking der Dokumente herangezogen. Das bedeutet, dass es für eine Anfrage nicht nur ein Ergebnis, sondern für verschiedene Teilmengen des Ergebnisses verschiedene Qualitätsmaße ermittelt werden. Dadurch wird dem Umstand Rechnung getragen, dass Benutzer bei gerankten Ergebnismengen – vor allem, wenn diese sehr groß sind – nicht die gesamte Menge, sondern nur die ersten Treffer ansehen, bis sie ihr Informationsbedürfnis befriedigt sehen oder keine relevanten Dokumente mehr erwarten.

In dem Werkzeug für die Relevanzbewertung werden einige wichtige Relevanzstufen direkt angezeigt (siehe Abb. 3). Die einzelnen Recall- und Precision-Werte werden dabei effektiv im Zusammenspiel zwischen Datenbankindices und Java-Anwendung berechnet. Eine Export-Funktion erlaubt die Berechnung und das Abspeichern vollständiger Wertelisten für Recall-Precision-Graphen oder als Graphen mit Recall und Precision im Verhältnis zu den Cut-Off-Werten (siehe Abb. 4).

Eine weitere Auswertung der exportierten Recall-Precision-Werte mit Hilfe von Werkzeugen wie Microsoft Excel, SPSS oder auch IR-STAT-PAK (Blustein & Tague-Sutcliffe 1995) ist erforderlich. Angesichts der nach wie vor geringen Zahl von Testanfragen sind statistisch signifikante Ergebnisse allerdings kaum zu erwarten. (siehe auch Braschler 2002:19ff)

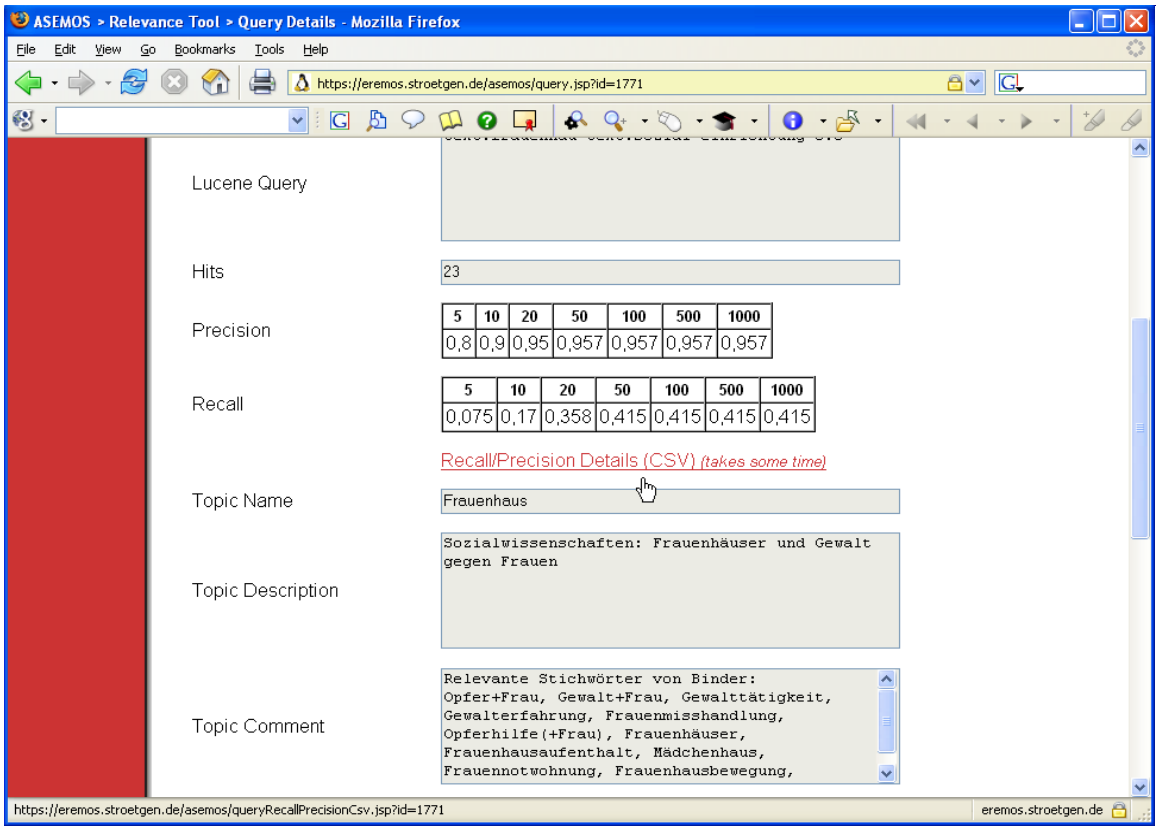

Abbildung 3: Recall und Precision im ASEMOS Relevance Tool

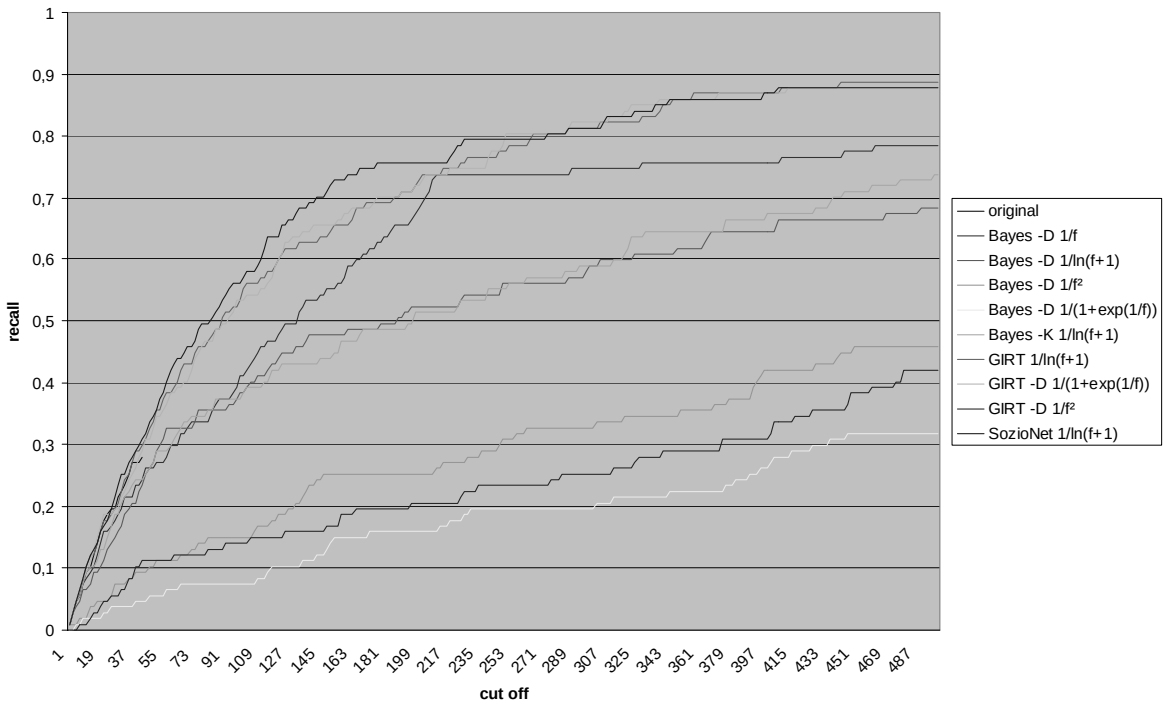

Abbildung 4: Beispielgraph mit Recall-Werten für einige Anfragen zum Thema „Arbeitsszeit und Flexibilität

3 Ergebnisse und Ausblick

Die Auswirkung von Anfragetransfers auf der Basis statistisch erzeugter semantischer Relationen ist nach wie vor unzureichend evaluiert. Die Arbeiten im Projekt *ASEMOS* sollen einen Beitrag dazu leisten, in dieser Lücke besser begründete Urteile fällen zu können.

Erste Auswertungen deuten darauf hin, dass die Anwendung von Anfragetransfers auf der Basis statistisch erzeugter semantischer Relationen sowohl in Bezug auf den Recall als auch die Precision eine teilweise deutliche Verbesserung der Retrievalqualität bewirken kann. Eine systematische Auswertung steht nun an.

Weiterhin werden die beschriebenen Verfahren auf den Bereich Patentintformation übertragen, dort finden natürlich entsprechende Anstrengungen zur Evaluierung statt.[15]

Der immense Aufwand bei der Erstellung von Relevanzurteilen erlaubt leider nur kleine Fortschritte. Die bisher am IZ Sozialwissenschaften und der Universität Hildesheim geleisteten Arbeiten lassen sich aber weiter verwerten und ausbauen: Testkorpora, Testanfragen und Relevanzurteile können für zukünftige Retrievaltests im Kontext Webretrieval für die Sozialwissenschaften eine wertvolle Ausgangsbasis liefern.

Literatur

Binder, Gisbert; *Marx*, Jutta; *Mutschke*, Peter; *Strötgen*, Robert; *Plümer*, Judith; *Kokkelink*, Stefan (2002)
Heterogenitätsbehandlung bei textueller Information verschiedener Datentypen und Inhaltserschließungsverfahren. Bonn: IZ Sozialwissenschaften (IZ-Arbeitsbericht; Nr. 24).

Blustein, James; *Tague-Sutcliffe*, Jean
IR-STAT-PAK. Version: 1995. http://www.csd.uwo.ca/~jamie/IRSP-overview.html. – Online–Ressource, Abruf: 2005-03-02

Braschler, Martin (2002)
CLEF 2002. Overview of Results. In: *Peters*, Carol; *Braschler*, Martin; *Gonzalo*, Julio; *Kluck*, Michael (Hrsg.): Advances in Cross-Language Information Retrieval. Third Workshop of the Cross Language Evaluation Forum, CLEF 2002, Rome, Italy, September 19 - 20, 2002. Berlin, Heidelberg, New York : Springer (Lecture Notes in Computer Science; Bd. 2785). S. 9–27.

Hellweg, Heiko; *Krause*, Jürgen; *Mandl*, Thomas; *Marx*, Jutta; *Müller*, Matthias N.O.; *Mutschke*, Peter; *Strötgen*, Robert (2001)
Treatment of Semantic Heterogeneity in Information Retrieval (IZ-Arbeitsbericht; Nr. 23). Bonn.

Kölle, Ralph; *Mandl*, Thomas; *Schneider*, René; *Strötgen*, Robert (2004)
Weiterentwicklung des virtuellen Bibliotheksregals MyShelf mit Semantic Web-Technologie: Erste Erfahrungen mit informationswissenschaftlichen Inhalten. In: Ockenfeld, Marlies (Hrsg.): Information Professional 2011; 26. Online-Tagung der DGI, DGI, Frankfurt am Main, 15. bis 17. Juni 2004; Proceedings. Frankfurt am Main : DGI (Tagungen der Deutschen Gesellschaft für Informationswissenschaft und Informationspraxis; Bd. 7), S. 111–124.

Krause, Jürgen
Standardisierung von der Heterogenität her denken. Zum Entwicklungsstand bilateraler Transferkomponenten für digitale Fachbibliotheken. Bonn: IZ Sozialwissenschaften (IZ-Arbeitsbericht; Nr. 28).

Mandl, Thomas; *Womser-Hacker*, Christa (2002)
Virtual ontologies for browsing interfaces in digital libraries. In: *Isaías*, Pedro (Hrsg.): Proceedings of the 2nd International Workshop on New Developments in Digital Libraries (NDDL 2002). In conjunction with the 4th International Conference On Enterprise Information Systems (ICEIS). April 2, 2002. Ciudad Real, Spanien. S. 39–50.

15. Dafür an dieser Stelle vielen Dank vor allem an Herrn Dr. Michael Schwantner und Herrn Dr. Claus-Dieter Siems vom FIZ Karlsruhe und für die Arbeit an den Relevanzurteilen Frau Carina Völpel von der Universität Hildesheim!

Stempfhuber, Maximilian (2004)
Infoconnex. Der Informationsverbund Pädagogik - Sozialwissenschaften - Psychologie. In: Sharing Knowledge. Scientific Communication. 9. Kongress der IuK-Initiative der Wissenschaftlichen Fachgesellschaften in Deutschland. Bonn : IZ Sozialwissenschaften (Tagungsberichte; Bd. 8).

Strötgen, Robert (2002)
Behandlung semantischer Heterogenität durch Metadatenextraktion und Anfragetransfer. In: *Womser-Hacker*, Christa; *Wolff*, Christian; *Hammwöhner*, Rainer (Hrsg.): Information und Mobilität: Optimierung und Vermeidung von Mobilität durch Information; Proceedings des 8. Internationalen Symposiums für Informationswissenschaft (ISI 2002). Konstanz : UVK (Schriften zur Informationswissenschaft; Bd. 40), S. 259–271.

Strötgen, Robert (2004)
ASEMOS. Weiterentwicklung der Behandlung semantischer Heterogenität. In: *Bekavac*, B.; *Herget*, J. & *Rittberger*, M. (Hrsg.): Information zwischen Kultur und Marktwissenschaft; Proceedings des 9. Internationalen Symposiums für Informationswissenschaft (ISI 2004). Konstanz : UVK, 2004 (Schriften zur Informationswissenschaft; Bd. 42), S. 269–281.

Strötgen, Robert; *Kokkelink*, Stefan (2001)
Metadatenextraktion aus Internetquellen: Heterogenitätsbehandlung im Projekt CARMEN. In: Information research & content management; 23. Online-Tagung der DGI und 53. Jahrestagung der Deutschen Gesellschaft für Informationswissenschaft und Informationspraxis e.V., DGI, Frankfurt am Main, 8. bis 10. Mai 2001. Frankfurt am Main : DGI. S. 56–66.

Womser-Hacker, Christa (1989)
Der PADOK-Retrievaltest. Zur Methode und Verwendung statistischer Verfahren bei der Bewertung von Information-Retrieval-Systemen. Hildesheim and Zürich and New York : Georg Olms Verlag (Sprache und Computer; Bd. 10).