

Anfragetransfers zur Integration von Internetquellen in Digitalen Bibliotheken auf der Grundlage statistischer Termrelationen

Robert Strötgen

robert.stroetgen@tib.uni-hannover.de

TIB/UB Hannover

Welfengarten 1B

30167 Hannover

1 Kontext: Behandlung semantischer Heterogenität in Digitalen Bibliotheken

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 (Schott/Schroeder 2004) und statistisch erzeugte Relationen angegangen. (Hellweg et al. 2001) Ein Teil der Ergebnisse der Arbeiten am IZ Sozialwissenschaften⁵ waren statistische Relationen zwischen Deskriptoren, die mittels Kookurrenzbeziehungen berechnet wurden. Diese Relationen wurden dann für die

¹ <http://dublincore.org/>

² <http://www.w3.org/RDF/>

³ <http://www.w3.org/2001/sw/>

⁴ Sonderfördermaßnahme im Rahmen von Global-Info (Content Analysis, Retrieval and MetaData: Effective Networking), <http://www.mathematik.uni-osnabrueck.de/projects/carmen/>

⁵ <http://www.gesis.org/iz/>

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.

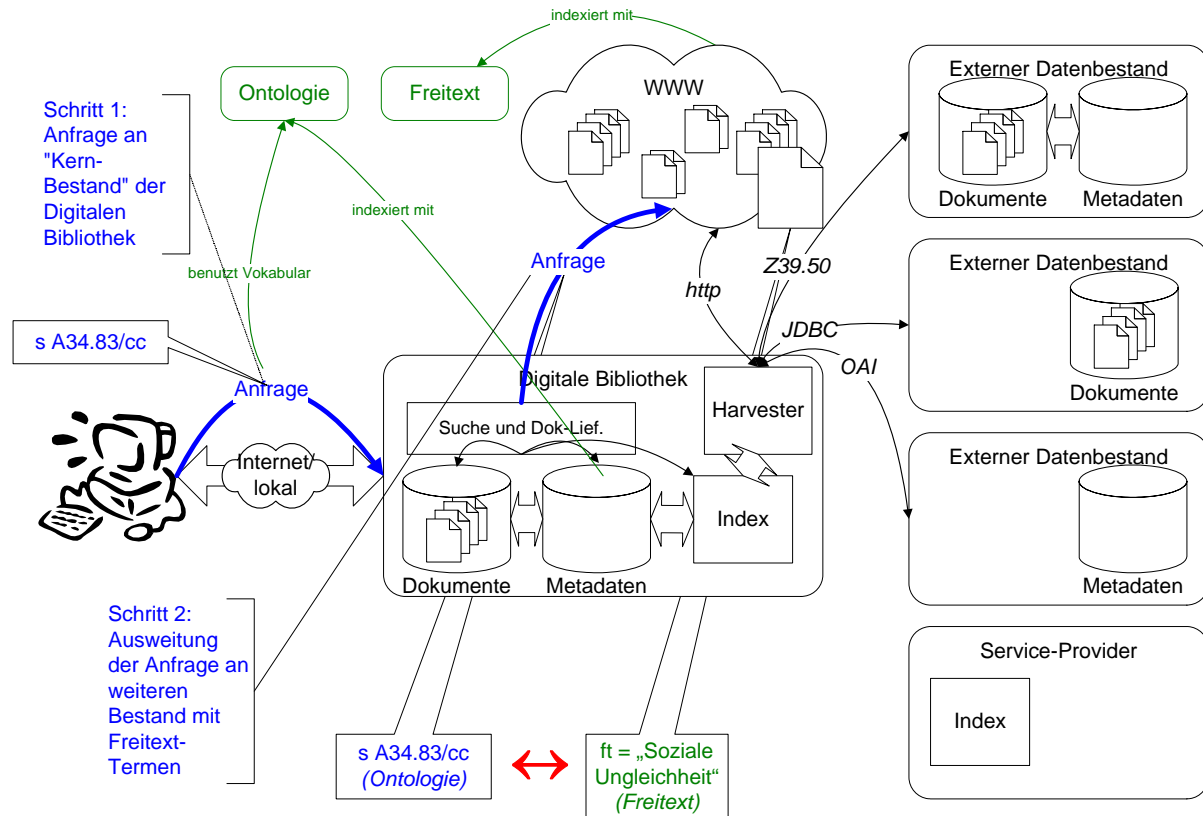


Abbildung 1: Beispiel-Szenario für semantische Heterogenität in digitalen Bibliotheken

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 einige Ansatzpunkte aufgezeigt, wie sich das Verfahren noch verbessern lassen könnte. An dieser Stelle setzt das Projekt ASEMOS⁶ an, in dem die CARMEN in entwickelten Ansätze weiter verfolgt werden. Verbesserungen wurden insbesondere durch ausgefeiltere 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 CARMEN zunehmend verdeutlicht. Integrierte Portale im Kontext digitaler Bibliotheken wie z.B. die Wissenschaftsportale Infoconnex⁷ und Vascoda⁸ planen den Einsatz entsprechender Anfrageübersetzungstechniken oder haben sie bereits eingeführt. (Stempfhuber 2004:243; Walter/Mayr/Stempfhuber/Ballay 2006) Auch im Kontext des Semantic Web wird die Vermittlung zwischen verschiedenen Ontologien als ein zentrales Problem der Integration

⁶ Applied Statistics, Evaluation and Machine Learning: Ontology Switching

⁷ <http://www.infoconnex.de/>

⁸ <http://www.vascoda.de/>

heterogener Dokumente angesehen und bearbeitet. Im näheren Kontext dieser Arbeit wird an der Universität Hildesheim bei dem virtuellen Bibliotheksregal *MyShelf* eine Vermittlung zwischen Klassifikationen durch Ontology Switching angewendet. (Mandl & Womser-Hacker 2002, Kölle et al. 2004)

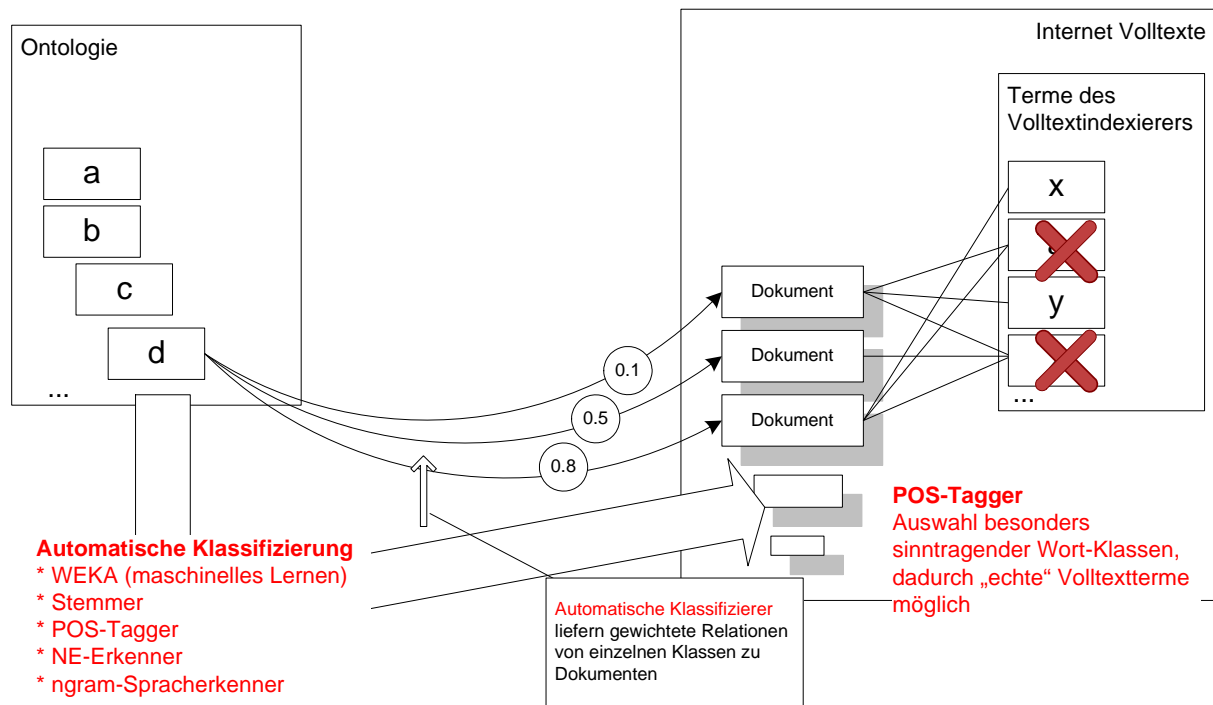


Abbildung 2: Parallelkorpussimulation in ASEMOS.

Auf das Verfahren zur statistischen Erstellung semantischer Relationen soll hier nicht genauer eingegangen werden. (siehe dazu Strötgen 2004) Die Weiterentwicklung des in *CARMEN* entwickelten Verfahrens in *ASEMOS* betrifft vor allem die Simulation von Parallelkorpora (siehe Abbildung 2), die Ausgangsbasis der Kookkurrenzanalyse sind. Auf Seiten der Zuordnung der Ontologie wurden linguistische Verfahren wie Part-of-Speech-Tagging, n-gramm-basierte Spracherkennung und Named-Entity-Erkennung sowie automatisches Klassifizieren mit der Bibliothek für maschinelles Lernen *WEKA*⁹ eingesetzt. Bei der Zuordnung von Freitexttermen wurde die Anzahl der Terme durch POS-Tagging so weit reduziert, dass hier im Unterschied zu *CARMEN* echte Freitextterme Verwendung finden konnten.

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.

⁹ <http://www.cs.waikato.ac.nz/ml/weka/>

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.¹⁰

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*¹¹ ein speziell angepasster Webroboter implementiert. Dieser wurde auf 11 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*¹² syntaktisch bereinigt und in eine *PostgreSQL*¹³-Datenbank abgespeichert.

Für die Domäne Patentdokumentation, auf die der Ansatz übertragen wurde, wurden Abzüge von Fachdatenbanken genutzt, die vom FIZ Karlsruhe zur Verfügung gestellt wurden.¹⁴ Es wurden zeitliche und thematische Ausschnitte aus folgenden Datenbanken genutzt: PATDPA, LPATDPA, EPFULL und LINSPEC. Dabei wurden die Korpora LPATDPA (29523 Dokumente) und EPFULL (10489 Dokumente) für das Lernen der semantischen Relationen und die Korpora LPATDPA (7744 Dokumente) und LINSPEC (11090 Dokumente) für die Evaluierung genutzt.

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 Anfragetransfers 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 13 Anfragen mehr als verdoppelt. Auch wenn eine deutlich höhere Anzahl an Testanfragen wünschenswert wäre, ist der damit verbundene Aufwand bei der Erstellung von Relevanzurteilen (dazu im Folgenden mehr) hier leider nicht zu leisten.

¹⁰ Eine genauere Beschreibung des Korpus findet sich in Binder et al. 2002:5-6.

¹¹ <http://www.matuschek.net/software/job/>

¹² <http://tidy.sourceforge.net/>

¹³ <http://www.postgresql.org/>

¹⁴ Dafür an dieser Stelle vielen Dank vor allem an Herrn Dr. Michael Schwantner vom FIZ Karlsruhe.

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*.¹⁵ Um eine Vergleichbarkeit mit dem 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 an der Universität Hildesheim und am IZ Sozialwissenschaften intellektuell geleistet.¹⁶

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.

Abbildung 3: Relevanzbewertung mit dem ASEMOS Relevance Tool

¹⁵ <http://lucene.apache.org/>

¹⁶ An dieser Stelle sei vor allem Frau Susanne Rauch und Herrn Christian Kolle vom IZ Sozialwissenschaften sowie Frau Kathrin Wünnemann, Herrn Robert Stoldt, Herrn Ben Heuwing und Frau Uta Fröhlich von der Universität Hildesheim für die unermüdliche Arbeit als Jurorinnen gedankt!

Zur Relevanzbewertung wird den Juroren eine Liste aller noch unbewerteten Dokumente zu einer Anfrage präsentiert. Die Reihenfolge der Dokumente ist dabei über eine Zufallsfunktion gemischt worden, um zu verhindern, dass die Juroren am Ende der Liste nicht relevante Dokumente erwarten und diese Erwartung 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 die Bereich Sozialwissenschaften und Patente wurden bislang zu insgesamt 1942 von 2145 Anfragevarianten der 86 Ausgangsfragen zu 23 Themenbereichen 114359 einzelne Relevanzurteile für die Dokumente in den Testkorpora erstellt. Selbst wenn die geübten Juroren je nach Dokument teilweise mehrere Relevanzurteile in einer Minute treffen können, ist der Aufwand immens.

2.4 Auswertung

Der Evaluierung wird das Szenario zu Grunde gelegt, dass ein Benutzer seine Anfrage spezifisch für eine Datenquelle formuliert. In einem zweiten Schritt wird die Anfrage auf eine zweite Datenquelle übertragen. Dabei wird zwischen zwei Varianten unterschieden: Einmal wird die Anfrage „naiv“ unverändert an die zweite Datenbasis gestellt, in der zweiten Variante wird sie spezifisch in das Vokabular der zweiten Datenbasis transferiert bzw. übersetzt (siehe Abbildung 4). Im Fall eines Übergangs von einer Anfrage an eine Fachdatenbank, die unter Verwendung eines Thesaurus formuliert wurde, wird die für Internetquellen transferierte Anfrage um zusätzliche Freitextterme erweitert.

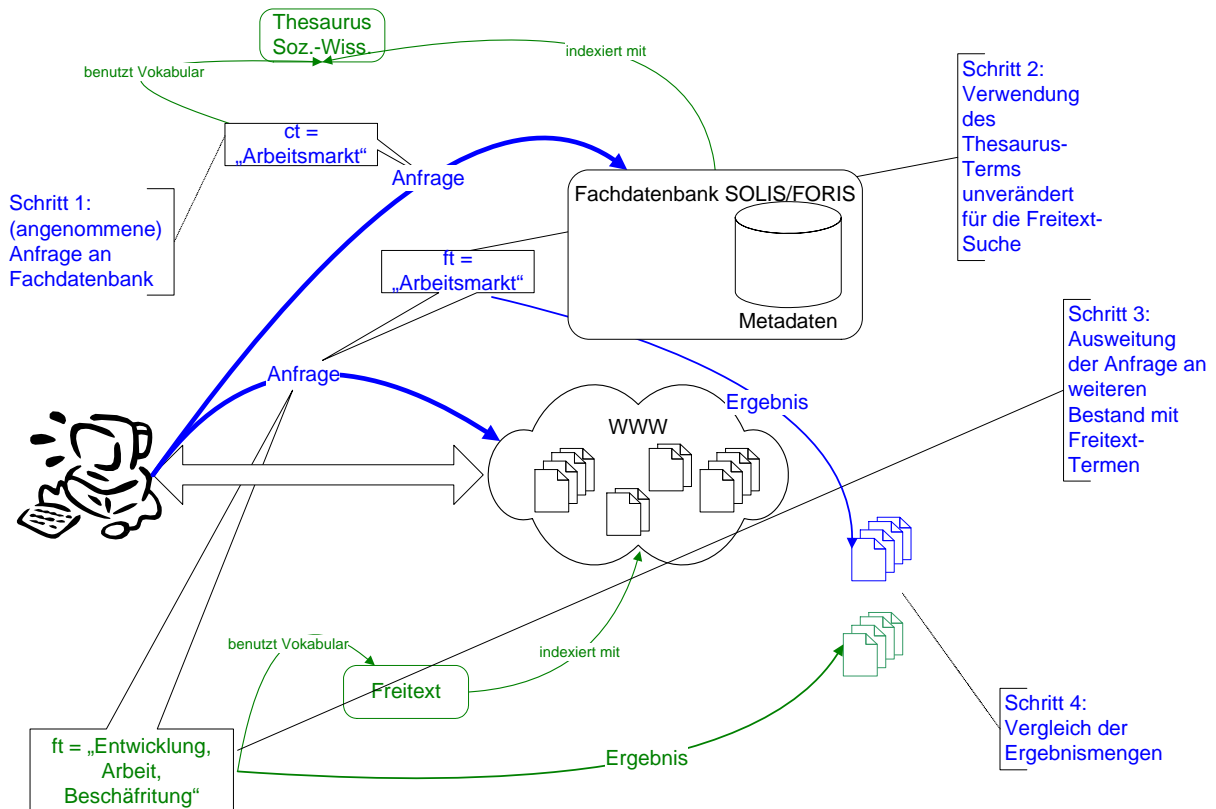


Abbildung 4: Szenario für die Evaluierung

Aber auch die Gegenrichtung wird einer Überprüfung unterzogen: Eine Freitext-Anfrage wird um geeignete Deskriptoren des Thesaurus ergänzt, bevor sie an eine Fachdatenbank gestellt wird. In beiden Fällen werden schließlich die Ergebnismengen verglichen und mit üblichen Maßen zur Retrievalqualität, v.a. Recall und Precision, bewertet.

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.

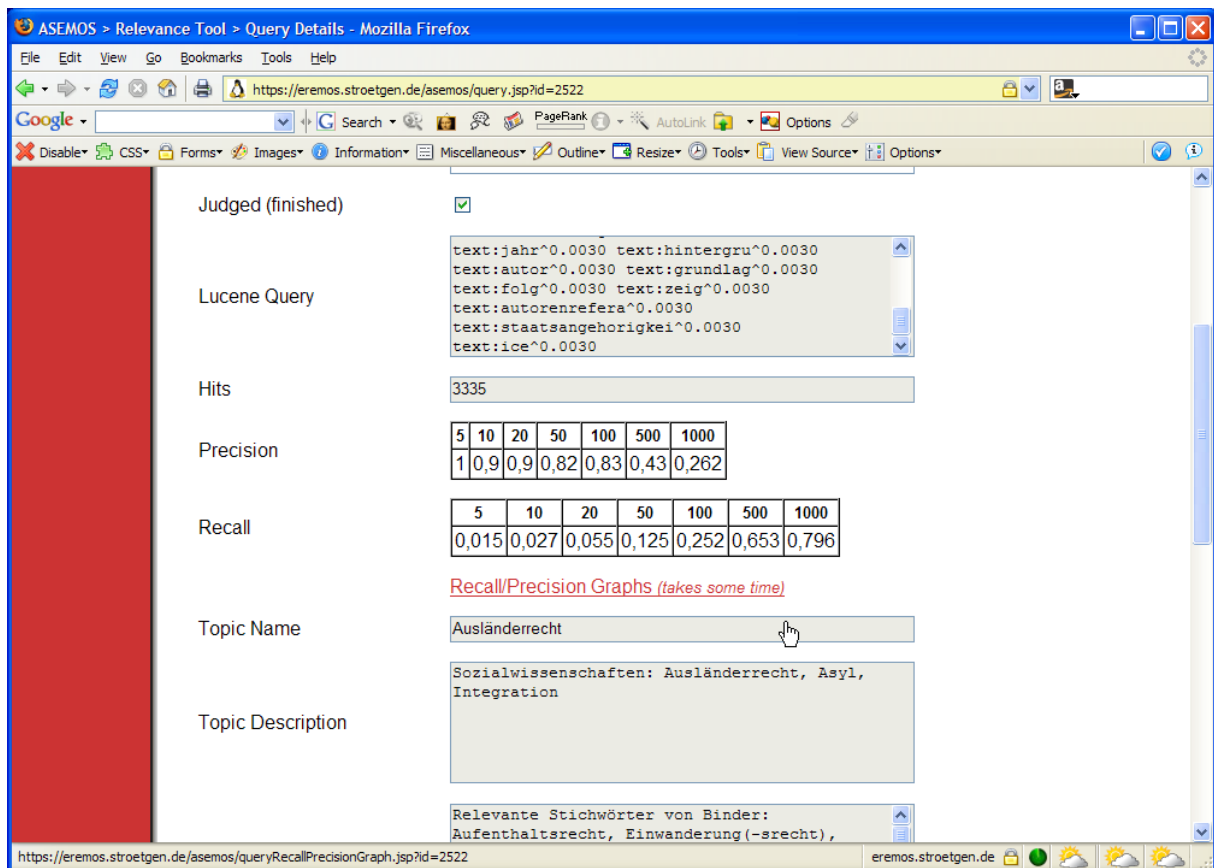


Abbildung 5: Recall und Precision im ASEMOS Relevance Tool

In dem Werkzeug für die Relevanzbewertung werden einige wichtige Relevanzstufen direkt angezeigt (siehe Abbildung 5). Die einzelnen Recall- und Precision-Werte werden dabei effektiv im Zusammenspiel zwischen Datenbankindices und Java-Anwendung berechnet. Es werden verschiedene Graphen erzeugt, z.B. Recall-Precision-Graphen (siehe Abbildung 6).

Darüber hinaus wurden in dem Relevanz-Werkzeug eine Reihe von statistischen Auswertungsverfahren implementiert, u.a. T-Test. Angesichts der nach wie vor geringen Zahl von Testanfragen lassen sich statistisch signifikante Ergebnisse allerdings kaum erzielen. (siehe auch Braschler 2002:19ff)

3 Ergebnisse

In einem ersten Beispiel wird eine Anfrage von einer sozialwissenschaftlichen Fachdatenbank für eine folgende Internetrecherche erweitert. Ausgangspunkt ist der Deskriptor „Fremdbild“

aus dem Thesaurus Sozialwissenschaften. Eine transferierte Anfrage lautet „fremdbild Sie (0.043) fh-jena (0.043) Jena (0.028) FH (0.028) Berlin (0.015) Matthias (0.015) Silke (0.015) Ketzler (0.015) Lawitschka (0.015) aktuell (0.014) juni (0.014) veranstaltung (0.014) uhr (0.014) raum (0.014) Deutschland (0.014) es (0.014) B (0.014) Marcel (0.014)“. Die Terme sind dabei teilweise bereits einer linguistischen Vorbehandlung unterzogen worden, in den Klammern finden sich spezifische Gewichte. Auf die Diskrepanz zwischen statistisch erzeugter semantischer Relationen und semantischer Evidenz wurde beispielsweise in Strötgen 2002 hingewiesen.

Wird die Ausgangsanfrage unverändert an die Internetdokumente geschickt, erhält der Benutzer ein Ergebnis von vier Treffern, von denen zwei relevant sind. Die transferierten Anfragen liefern wesentlich größere Treffermengen, im oben genannten Beispiel 3935, davon mehr als 80 relevante. Der Zugewinn an relevanten Treffern ist also erheblich. Wichtig für den Benutzer ist aber besonders, wie gut das Ranking in der großen Treffermenge die relevanten Dokumente nach vorne bewegt.

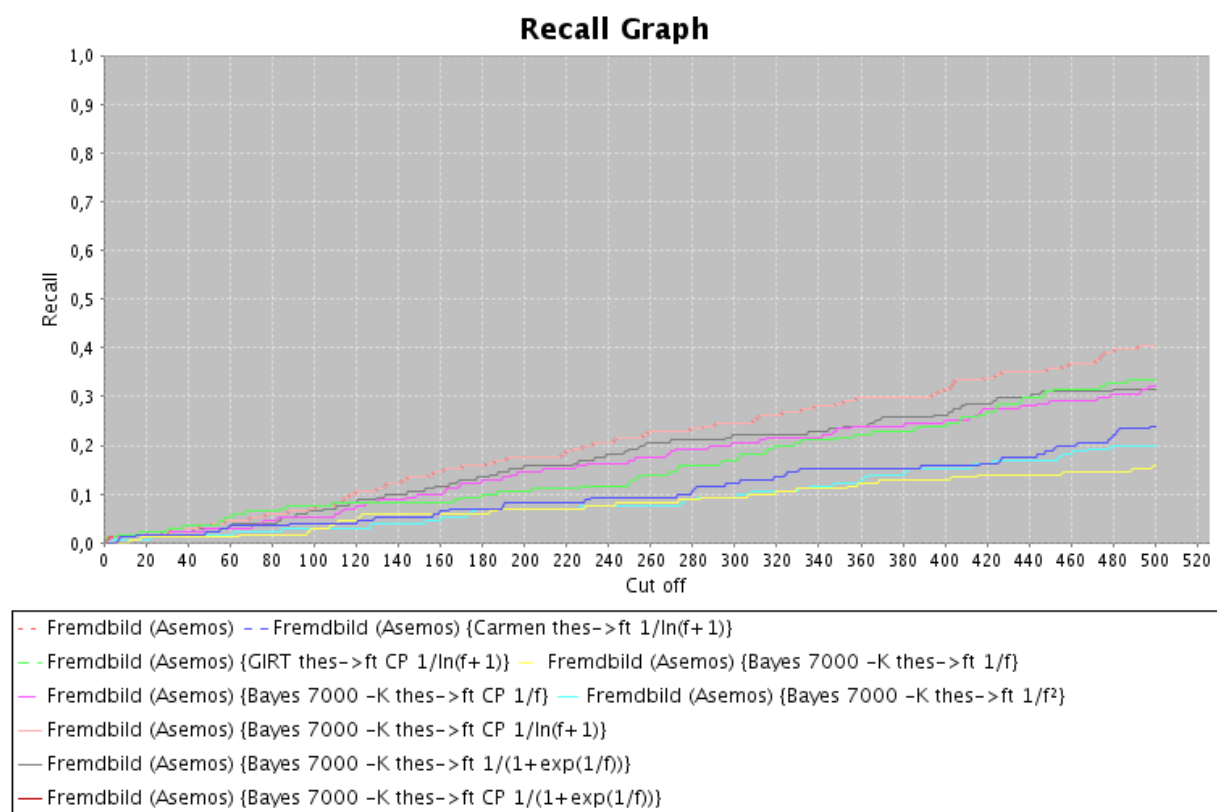


Abbildung 6: Recall für fortlaufende Schwellenwerte für einige Anfragen zum Thema „Fremdbild“

Abbildung 6 zeigt für die ersten 500 Dokumente, dass sich bei den transferierten Anfragen ein konstanter und mehr oder minder steiler Zugewinn an relevanten Treffern über die Treffermenge hinweg erzielen lässt. Der Trend der Original-Anfrage setzt sich dabei einigermaßen konstant fort. Weiterhin zeigt sich, dass der Zugewinn bei den in ASEMOS verbesserten Methoden nahezu doppelt so erfolgreich ist wie das in CARMEN entwickelte Verfahren.

Der Recall-Precision-Graph in Abbildung 7 zeigt, dass die untransferierte Ausgangsanfrage anfangs eine deutlich bessere Precision aufweist, bei zunehmendem Recall aber deutlich gegenüber den transferierten Anfragen zurückfällt. Die Original-Anfrage beginnt mit einer relativ guten Precision von 0,4 für den Recall 0 und fällt dann steil ab. Die transferierten Anfragen halten eine Precision von ca. 0,15 bis hin zu einem Recall von 0,4. Und auch hier

zeigt sich eine deutliche Verbesserung gegenüber der in *CARMEN* erreichten Retrievalqualität.

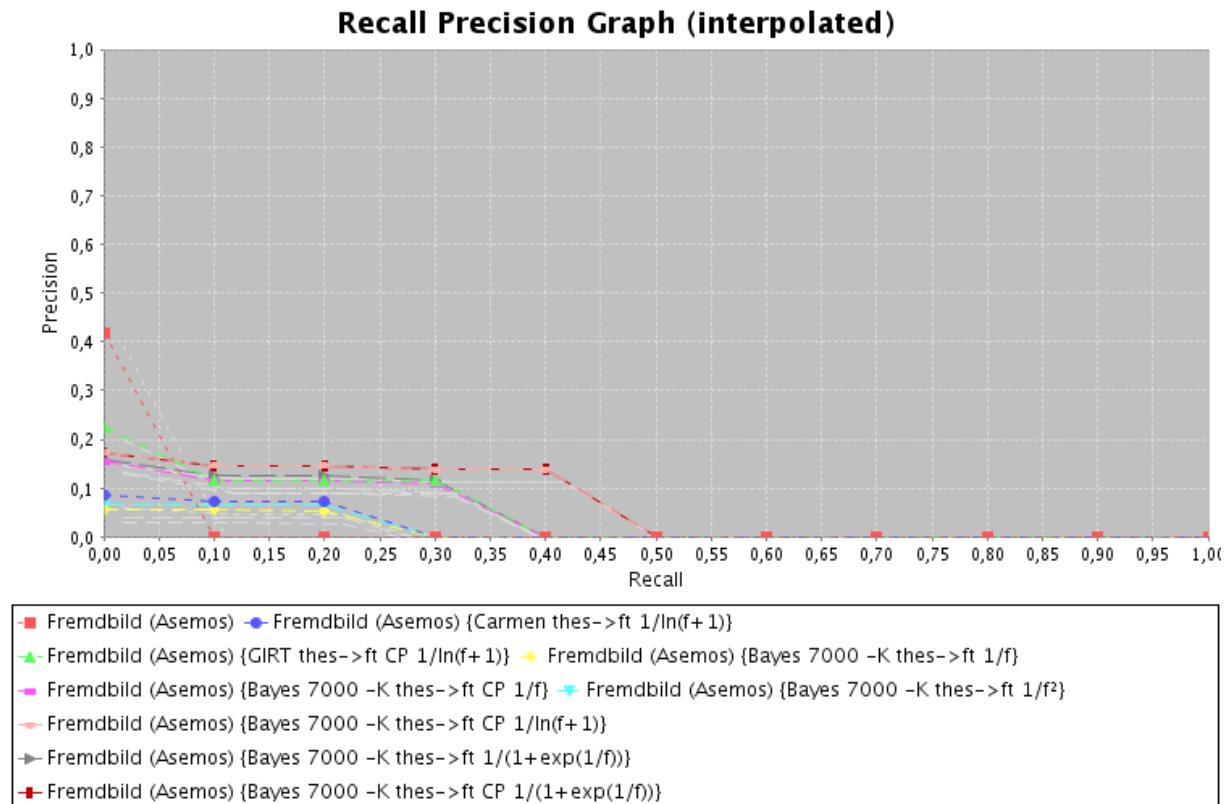


Abbildung 7: Recall-Precision-Werte für einige Anfragen zum Thema „Fremdbild“

Auch bei der Evaluierung der Gegenrichtung zeigt sich ein ähnliches Bild. Ein Beispiel für den Überstieg von einer Freitextanfrage an Internet-Dokumente zu einer Fachdatenbank-Anfrage ist die Anfrage nach „Arbeitszeit und Flexibilität“. Bei einem Transfer wird die Anfrage um die folgenden Terme aus dem Thesaurus Sozialwissenschaften erweitert (in Klammern wieder die jeweiligen Gewichte): „Qualitätssicherung (0.041) Arbeitsorganisation (0.041) Bildungsreform (0.041) Beschäftigung (0.041) Betreuung (0.041) Krise (0.041) Dienstleistung (0.041) Finanzierung (0.041) Interessenvertretung (0.041) Kind (0.041) Kompetenz (0.041) Leistung (0.041) Reform (0.041) Organisationsform (0.041) Sicherheit (0.041) audiovisuelle Medien (0.041) Lehrplan (0.041) tertiärer Sektor (0.041) Stabilität (0.041) Subvention (0.041)“.

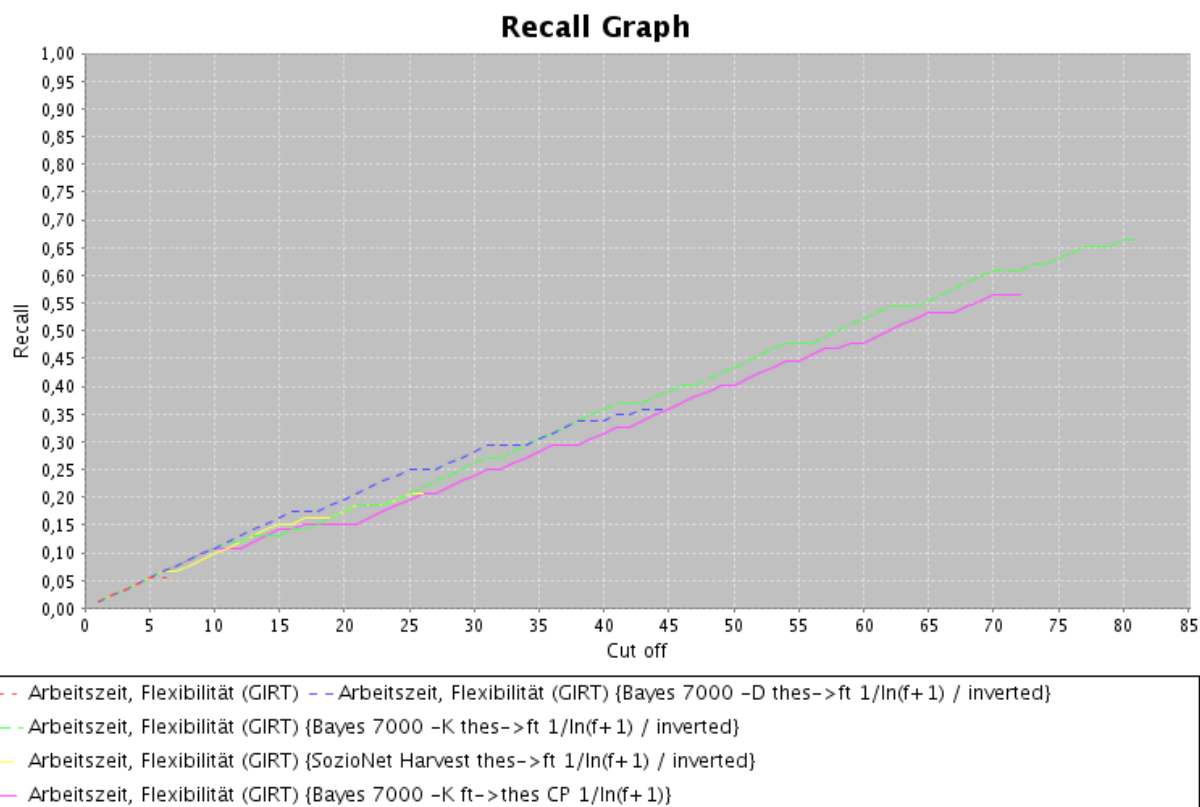


Abbildung 8: Recall für fortlaufende Schwellenwerte für einige Anfragen zum Thema „Arbeitszeit, Flexibilität“

Die Original-Anfrage an die GIRT-Datenbasis führt zu sieben Treffern, von denen fünf relevant sind. Die transferierten Anfragen führen auch hier zu deutlich mehr Treffern (für die Beispielanfrage 81 Treffer, von denen 53 relevant sind). Abbildung 8 zeigt auch hier den konstanten Zugewinn an relevanten Treffern gegenüber der untransferierten Anfrage. Der Recall-Precision-Graph in Abbildung 9 zeigt, dass die Ursprungsanfrage bei einem Recall von 0,2 abbricht. Die transferierten Anfragen weisen nicht nur einen deutlich besseren Recall auf, sondern führen auch zu einer gegenüber der untransferierten Anfrage wesentlich besseren Precision.

Auch wenn die Deskriptoren aus dem Thesaurus Sozialwissenschaften keineswegs einen evidenten Zusammenhang mit dem Thema der Anfrage haben, können mit ihnen trotzdem viele relevante Dokumente nachgewiesen werden.

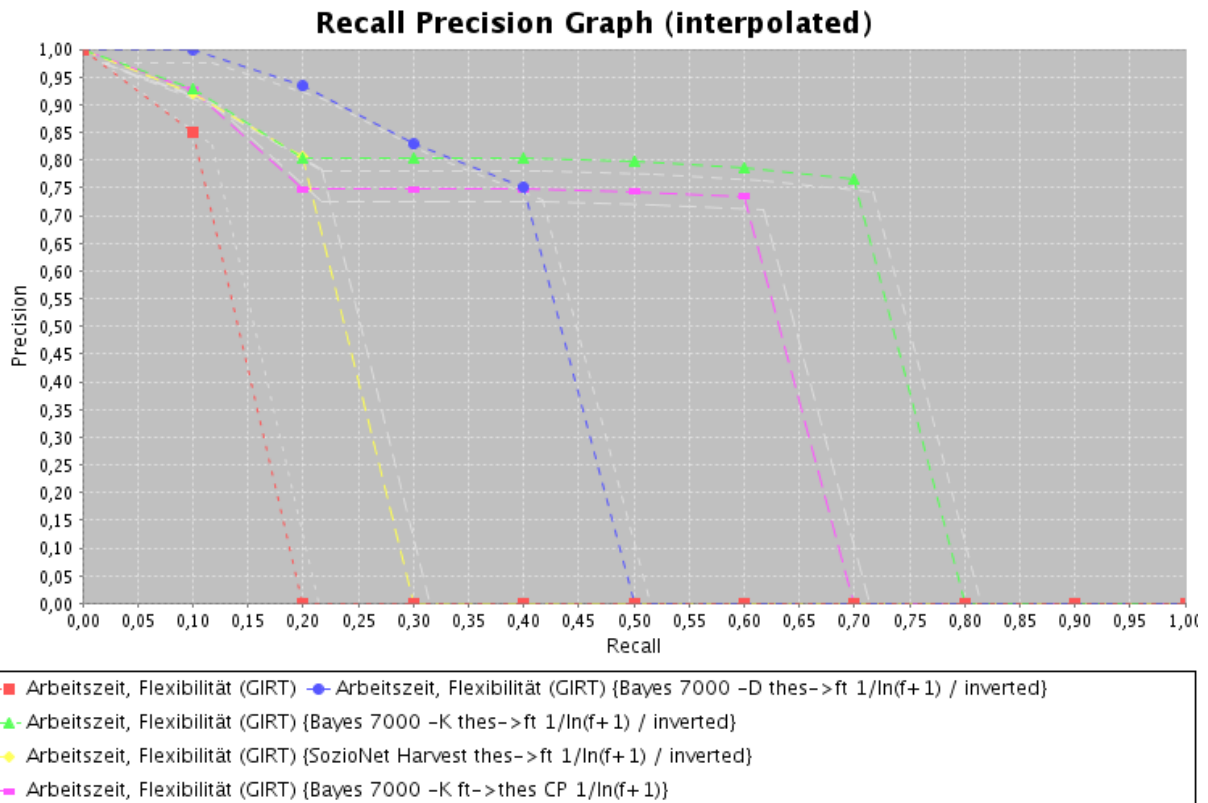


Abbildung 9: Recall-Precision-Werte für einige Anfragen zum Thema „Arbeitszeit, Flexibilität“

Bisher wurden nur einzelne Beispielanfragen angesprochen. Für eine Bewertung der Transfers ist aber ein statistischer Vergleich über die Menge aller Testanfragen unverzichtbar.

In Abbildung 10 zeigt der kombinierte Recall-Precision-Graph beispielhaft für die sozialwissenschaftlichen Internetdokumente die Mittelwerte für alle Original-Anfragen sowie für einige Transfers. Alle ausgewählten Transfer-Varianten liegen weitgehend oberhalb der untransferierten Anfragen (außer den Transfers aus *CARMEN*, die für niedrige Recall-Werte eine schlechtere Precision aufweisen). Wegen diverser Schnittpunkte der einzelnen Linien lässt sich aber schwer eine einheitliche Aussage treffen.

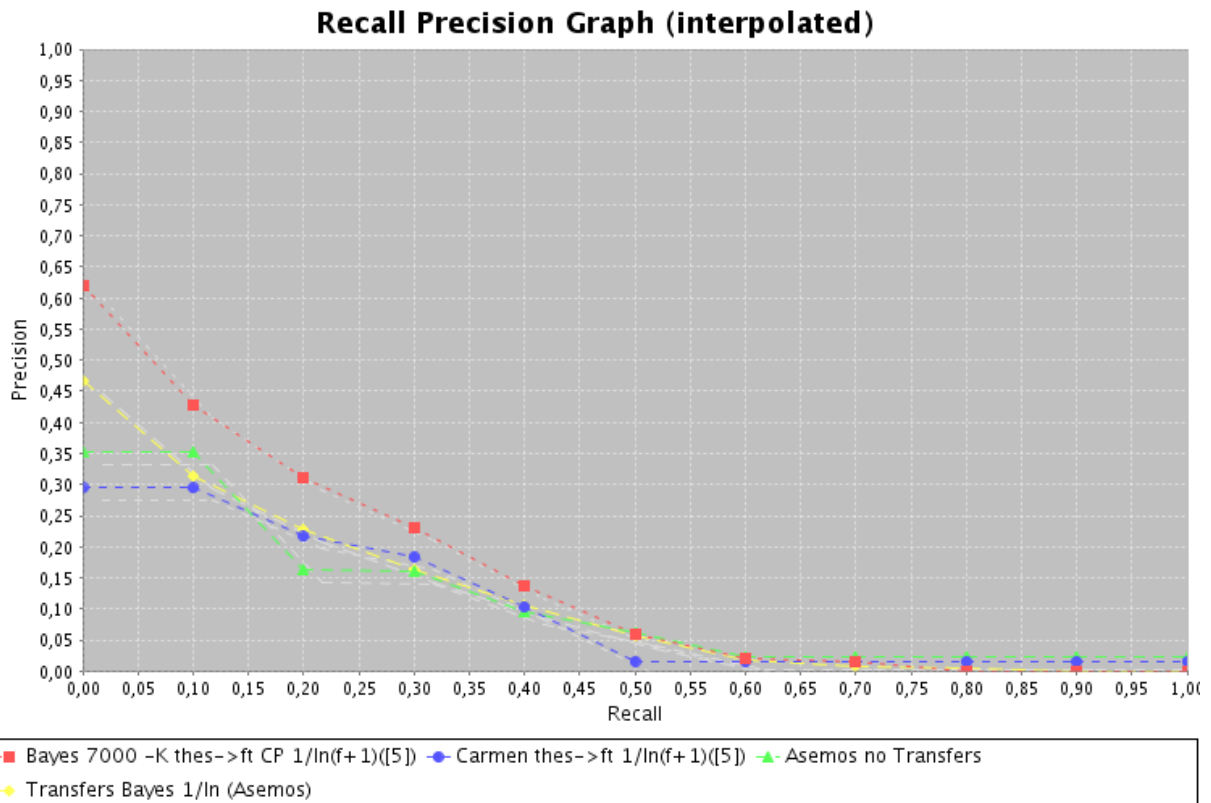


Abbildung 10: kombinierte Recall-Precision-Werte

Abbildung 11 zeigt im direkten Vergleich der Originalanfragen mit einem Transfertyp, wie Anfragen ohne Transfer überwiegend schlechtere Werte erzielen als Anfragen mit Transfer (rechte Seite). Schlechter ist der Mean Average Precision, was durch die größeren Treffermengen bei transferierten Anfragen erklärt werden könnte. Alle Vergleiche blieben allerdings unterhalb der statistischen Signifikanz. Dies ist angesichts der nach wie vor relativ geringen Anzahl an Testfällen nicht erstaunlich. Trotzdem dürfen sie wohl als Anzeichen dafür angesehen werden, dass auch über Einzelfälle hinaus generelle Vorteile für die Retrievalqualität durch die Anfragentransfers erzielt werden können.

Test	Asemos no Transfers			Bayes 7000 -K thes->ft 1/ln(f+1)([5])			t-Value	p-Value
	Mean	SD	Variance	Mean	SD	Variance		
Mean Average Precision	0,026735	0,0438933	0,0019266	0,0002881	0,0002035	0	2,1724118	0,0505748
Mean Reciprocal Rank	0,5833333	0,3818813	0,1458333	0,8352826	0,3483952	0,1213792	-1,6029681	0,1259966
Precision for Recall 0.0	0,336409	0,3495724	0,1222008	0,6186176	0,3434622	0,1179663	-1,88108	0,0766433
Precision for Recall 0.1	0,3522459	0,322826	0,1042166	0,4221086	0,2793507	0,0780368	-0,5408174	0,5949589
Precision for Recall 0.2	0,1643677	0,2855883	0,0815607	0,3051845	0,2680785	0,0718661	-1,1792609	0,2535868
Precision for Recall 0.3	0,1612264	0,2710239	0,0734539	0,2171756	0,2149118	0,0461871	-0,5388155	0,5961031
Precision for Recall 0.4	0,0954939	0,1875052	0,0351582	0,1202224	0,1553935	0,0241471	-0,3369025	0,7398394
Precision for Recall 0.5	0,0608271	0,1499984	0,0224995	0,0666079	0,1100799	0,0121176	-0,104211	0,9180459
Precision for Recall 0.6	0	0	0	0,0164638	0,0439378	0,0019305	-1,1241189	0,2935588
Precision for Recall 0.7	0	0	0	0,0018238	0,0054714	0,0000299	-1	0,3465935
Precision for Recall 0.8	0	0	0	0	0	0	./.	./.
Precision for Recall 0.9	0	0	0	0	0	0	./.	./.
Precision for Recall 1.0	0,0241792	0,0871794	0,0076002	0	0	0	1	0,3370491

Abbildung 11: Statistischer Vergleich von Anfragen ohne oder mit Transfer.

Die beschriebenen Verfahren wurden auf den Bereich Patentinformation übertragen und auf ähnliche Weise evaluiert.¹⁷ Das Szenario stellt sich hier allerdings ein wenig anders dar. Der Überstieg findet hier nicht zwischen Fachdatenbanken und Internetquellen, sondern zwischen verschiedenen Fachdatenbanken statt.

Im ersten Beispiel wird von einem Benutzer ausgegangen, der von einer Schlagwort-Suche in einer Ingenieursdatenbank (INSPEC) zu einer Patentrecherche übergeht. Thema der Recherche im Beispiel ist die Suche nach Elektronischen Programmführern („electronic program guide“). Die aus der Schlagwortsuche resultierende Freitextanfrage lautet: „epg (+interact* +(program* programm*) +(guid* assist*) +(tv televis*))“. Diese Freitextanfrage wird im zweiten Schritt um Klassen aus der Internationalen Patentklassifikation (IPC) erweitert. Über statistische Relationen werden in einem beispielhaften Transfer folgende IPC-Klassen zusätzlich in die Anfrage aufgenommen: H04N005-445 (0.236): Programmierung des Fernsehempfängers zur sichtbaren Wiedergabe zusätzlicher Information; H04N007-16 (0.059): Fernsehsysteme - Geheimsysteme; Abonnement-Systeme; H04N007-173 (0.052): Zweiwegbetrieb, z.B. sendet der Abonnent ein Programmwählsignal. In den Klammern finden sich wieder die jeweiligen Gewichte.

Abbildung 12 zeigt deutlich, dass auch in diesem Beispiel im Vergleich zur Ausgangsanfrage durch die Transfers ein deutlicher Zugewinn relevanter Dokumente erzielt werden kann. Während die Ausgangsanfrage bei 18 Treffern 17 relevante Patente nachweist, werden in dem im Beispiel genannten Transfer 1484 Treffer mit insgesamt 53 relevanten Patenten gefunden. Der Zugewinn an relevanten Patenten kann hier als besonders wichtig angesehen werden, da bei Patentrecherchen in sehr vielen Fällen auf Vollständigkeit hohen Wert gelegt wird.

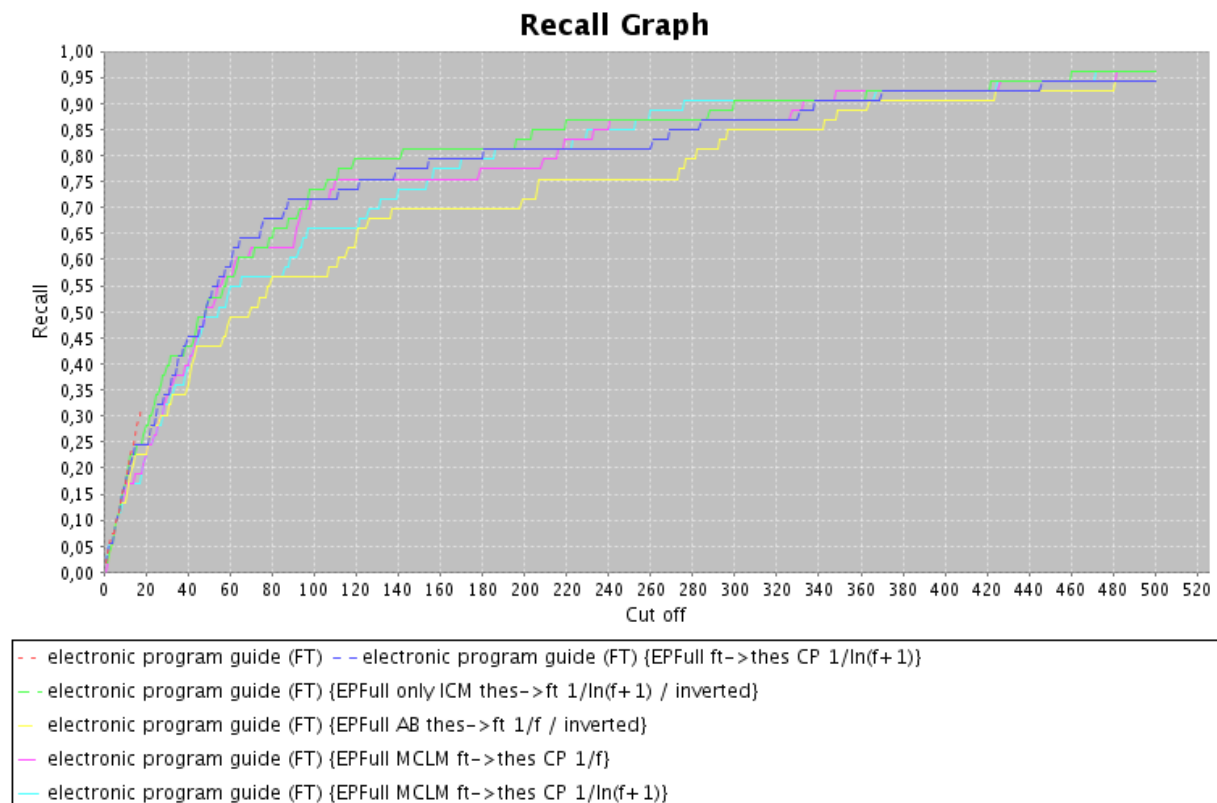


Abbildung 12: Recall für fortlaufende Schwellenwerte für einige Anfragen zum Thema „electronic program guide“

¹⁷ Dafür an dieser Stelle vielen Dank an Herrn Dr. Claus-Dieter Siems vom FIZ Karlsruhe und für die Arbeit an den Relevanzurteilen Frau Carina Vöpel und Ben Heuwing von der Universität Hildesheim!

Im Recall-Precision-Graph (Abbildung 13) lässt sich erkennen, dass die Ausgangsanfrage für niedrige Recall-Werte eine bessere Precision aufweist. Der Recall bricht allerdings bei 0,4 ein. Alle Anfragetransfers führen zu erheblich besserem Recall.

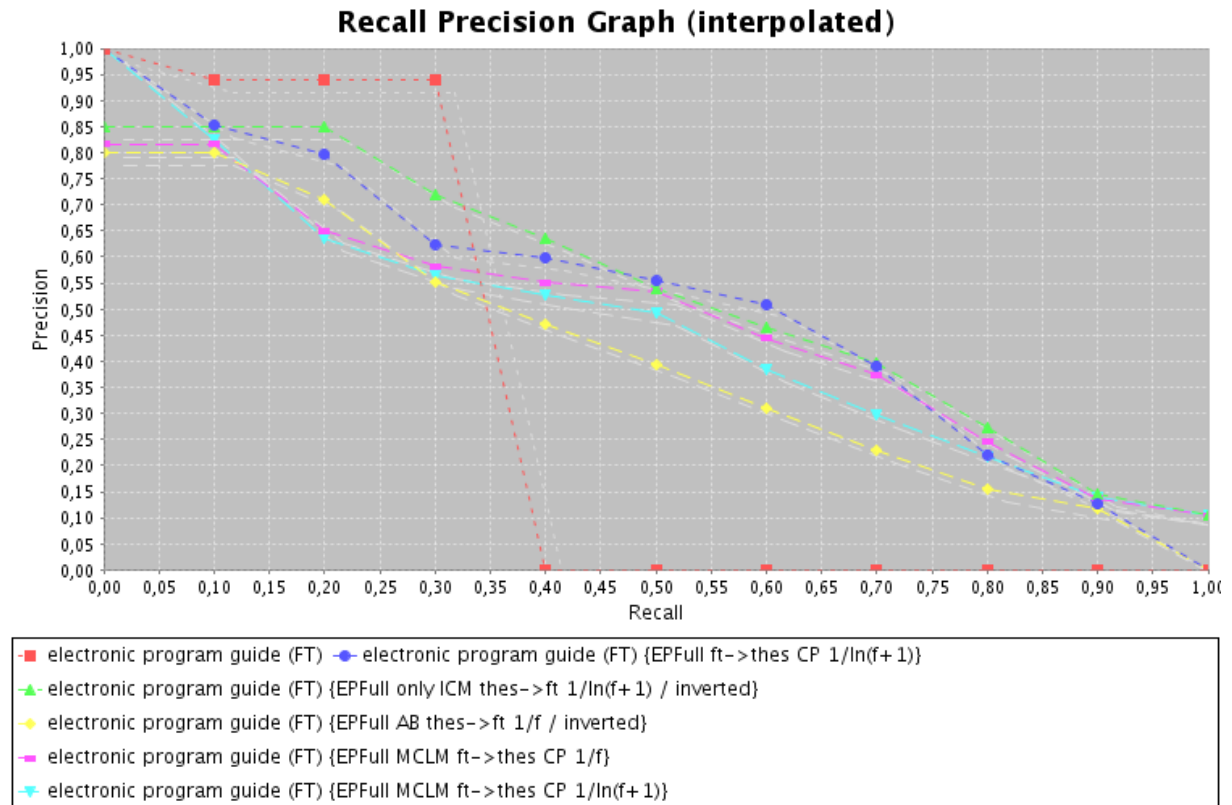


Abbildung 13: Recall-Precision-Werte für einige Anfragen zum Thema „electronic program guide“

Das zweite Szenario geht davon aus, dass eine Anfrage in einer Patentrecherche aus einer Mischung von Freitexttermen und IPC-Klassen besteht (die Formulierung der Anfragen erfolgte in Absprache mit Patentexperten aus dem FIZ Karlsruhe). Für den Themenbereich „Datenübertragung im Mobilfunk“ wurde folgende Ausgangsanfrage erstellt: „(+ (text:ra (+text:data* +(text:remot* text:transmiss* text:commun*))) +text:mobil*) ((ipc:H04Q007-00 ipc:H04Q007-20 ipc:H04Q007-22 ipc:H04Q007-38)^0.5)“.

Bei der Übersetzung von Anfragen wurden die IPC-Klassen, die in der Ingenieurdatenbank (INSPEC) keine Verwendung finden, durch zusätzliche Freitextterme ersetzt. Ein Beispiel für eine solche Übersetzung sieht so aus: „(+ (ra (+data* +(remot* transmiss* commun*))) +mobil*) (((system (0.106) method (0.104) syst (0.055) commun (0.055) termin (0.05) mobil (0.05) network (0.05) data (0.044) wireless (0.041) apparatu (0.041) messag (0.041) inform (0.041) wireless (0.041) dynam (0.033) set (0.033) number (0.033) updat (0.033) connect (0.033) predefin (0.033) handoff (0.033) multi (0.033) request (0.033) devic (0.033) creation (0.033) telepho (0.033)) (method (0.18) system (0.171) wireless (0.117) commun (0.1) network (0.067) data (0.061) mobil (0.061) multipl (0.058) control (0.058) devic (0.058) station (0.058) radio (0.058) apparatu (0.058) access (0.058) provid (0.058) (0.055) servic (0.055) compri (0.05) call (0.05) claim (0.05) wireless (0.05) messag (0.05) cellular (0.05) inform (0.05) signal (0.05)) (method (0.348) system (0.268) mobil (0.234) claim (0.201) commun (0.201) network (0.197) messag (0.194) servic (0.194) compri (0.188) data (0.184) abstract (0.175) draw (0.175) imag (0.175) receiv (0.159) radio (0.146) user (0.146) provid (0.146) second (0.138) termin (0.138) said (0.138) transmiss (0.138) step (0.134) control (0.125) inform (0.125) packet (0.125)) (method (0.796) system (0.662) commun (0.564) mobil (0.48) network (0.475) claim (0.472) compri (0.452) imag (0.395) draw (0.392) abstract (0.392) receiv

(0.375) base (0.368) station (0.365) control (0.355) transmit (0.348) radio (0.341) said (0.311) inform (0.308) further (0.305) wireless (0.298) provid (0.288) step (0.288) data (0.284) determin (0.284) signal (0.261)) (0.5)“

Die Ausgangsanfrage führt zu einer Treffermenge mit 9 Dokumenten, von denen nur eines relevant ist. Die im Beispiel beschriebene Transferanfrage führt zu insgesamt 7660 Treffern, von denen 16 relevant sind. Angesichts der Größe der Treffermenge ist das Ranking besonders wichtig. Innerhalb der ersten 20 Treffer finden sich immerhin schon 3, innerhalb der ersten 50 Treffer 6 relevante Dokumente. Auch hier lässt sich also ein Zugewinn an relevanten Dokumenten bei mäßigem Ballast antreffen. Der konstante Zugewinn an relevanten Dokumenten relativ schnell in der gerankten Treffermenge lässt sich auch in Abbildung 14 erkennen.

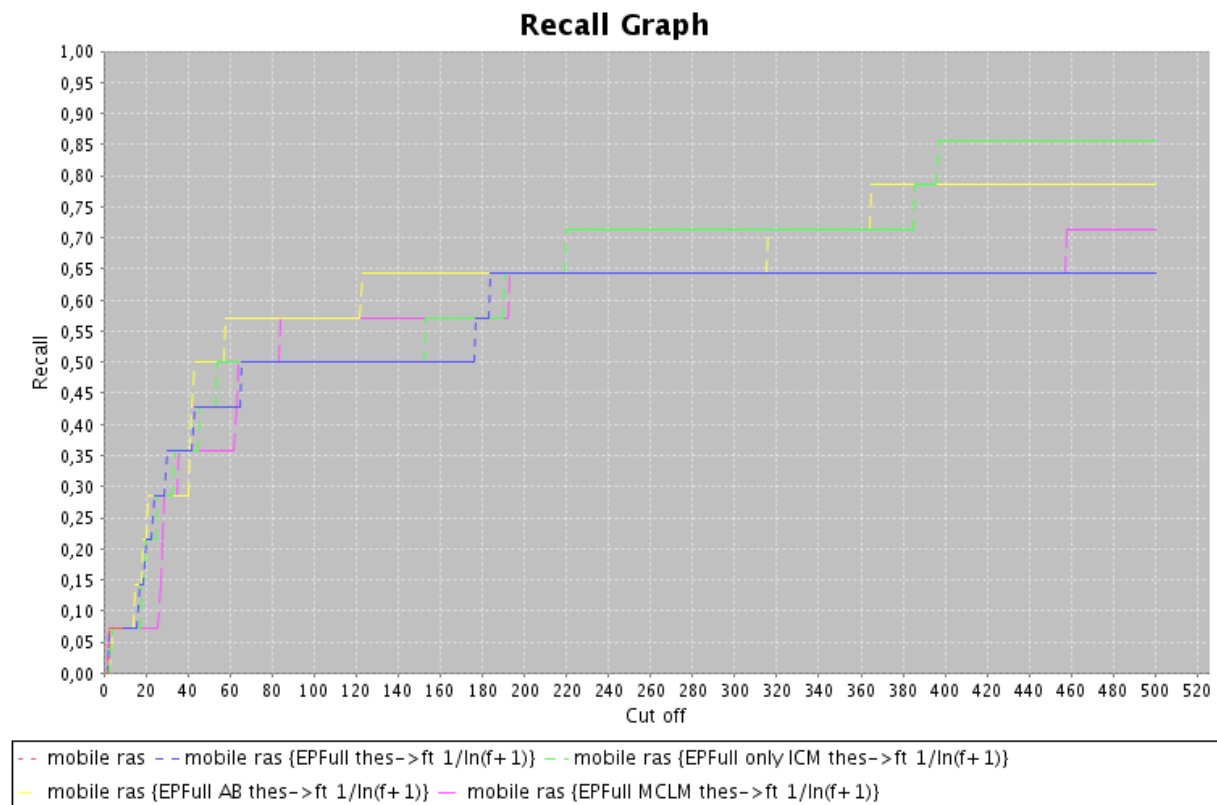


Abbildung 14: Recall für fortlaufende Schwellenwerte für einige Anfragen zum Thema „mobile ras“

Auch beim Recall-Precision-Graph (Abbildung 15) kann man sehen, dass die Ausgangsanfrage eine besseren Precision bei niedrigen Recall-Werten aufweist, allerdings beim Recall von allen anderen Anfragen deutlich übertroffen wird. Auch in diesem Beispiel ist der Zugewinn besonders wichtig, da mit der Ausgangsanfrage nur ein einziges relevantes Dokument gefunden werden konnte.

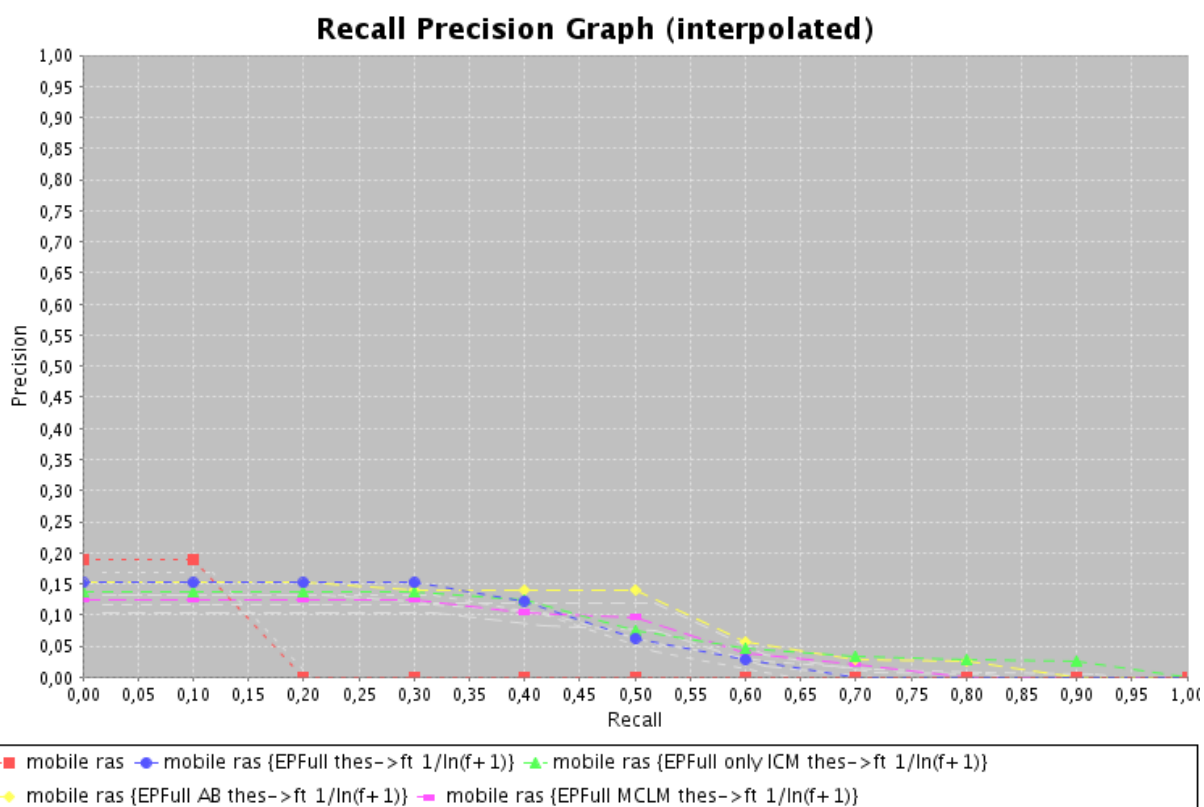


Abbildung 15: Recall-Precision-Werte für einige Anfragen zum Thema „mobile ras“

Bei der statistischen Betrachtung aller Anfragen für die Domäne Patente zeichnet sich ein ähnliches Bild wie für die Sozialwissenschaften ab: Transferierte Anfragen weisen zwar einen teilweise deutlich schlechteren Wert für die Mean Average Precision auf, können aber einen teilweise deutlich verbesserten Recall vorweisen.

Da in diesem Bereich aber die Erstellung der Relevanzurteile nach wie vor nicht abgeschlossen ist, sind an dieser Stelle leider keine weiteren statistischen Aussagen über alle Anfragen im Bereich möglich.

4 Zusammenfassung und Ausblick

Zusammenfassend für das Projekt *ASEMOS* lässt sich festhalten:

- Anfragetransfers liefern in aller Regel zusätzliche relevante Dokumente.
- Außer der (erwarteten) Verbesserung des Recall kann teilweise auch die Precision deutlich verbessert werden.
- Echte Parallelkorpora scheinen simulierten Parallelkorpora überlegen zu sein, aber auch diese liefern teilweise annähernd gute Ergebnisse.
- Die Parameter der Kookurrenzanalyse haben erheblichen Einfluss auf das Ergebnis.
- Die Ausnutzung der Gewichte verbessert die Ergebnisse erheblich.
- Gegenüber den Ergebnissen von *CARMEN* konnte bezüglich Recall und Precision ein deutlicher Zugewinn erzielt werden.
- Die Ergebnisse aus den Sozialwissenschaften können auf die Domäne Patente übertragen werden.
- Der Zugewinn an relevanten Dokumenten ist hier besonders relevant, da das Ziel in aller Regel ein sehr hoher Recall ist.

Die Analyse der Auswirkung von Anfragetransfers auf der Basis statistisch erzeugter semantischer Relationen erforderte weitere Anstrengungen. Die Arbeiten im Projekt *ASEMOS* haben einen Beitrag dazu geleistet, in dieser Lücke begründetere Urteile fällen zu können.

Testkorpora, Testanfragen und Relevanzurteile können für zukünftige Retrievaltests im Kontext Webretrieval für die Sozialwissenschaften eine wertvolle Ausgangsbasis liefern.

5 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).
- 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.
- Schott, Hannelore; Schroeder, Albert (2004): Crosskonkordanzen von Thesauri und Klassifikationen. In: Budin, Gerhard; Ohly, H. Peter (Hrsg.): *Wissensorganisation in kooperativen Lern- und Arbeitsumgebungen: Proceedings der 8. Tagung der Deutschen Sektion der Internationalen Gesellschaft für Wissensorganisation, Regensburg, 9.-11. Oktober 2002*. Würzburg: Ergon Verl. (Fortschritte in der Wissensorganisation; Bd. 8), S. 41 - 49
- 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.
- Walter, Anne-Kathrin; Mayr, Philipp; Stempfhuber, Maximilian; Ballay, Arne (2006): Crosskonkordanzen als Mittel der Heterogenitätsbehandlung in Informationssystemen. In: Stempfhuber, Maximilian (Hrsg.): *In die Zukunft publizieren: Herausforderungen an das Publizieren und die Informationsversorgung in den Wissenschaften; 11. Kongress der IuK-Initiative der Wissenschaftlichen Fachgesellschaften in Deutschland*. Bonn: Informationszentrum Sozialwissenschaften (Tagungsberichte; Bd. 11), S. 205 - 226.
- 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).