In: Hammwöhner, Rainer; Wolff, Christian; Womser-Hacker, Christa (Hg.): Information und Mobilität, Optimierung und Vermeidung von Mobilität durch Information. Proceedings des 8. Internationalen Symposiums für Informationswissenschaft (ISI 2002), Regensburg, 8. – 11. Oktober 2002. Konstanz: UVK Verlagsgesellschaft mbH, 2002. S. 259 – 271

Behandlung semantischer Heterogenität durch Metadatenextraktion und Anfragetransfers

Robert Strötgen

Informationszentrum Sozialwissenschaften
Lennéstr. 30
D-53113 Bonn
stroetgen@bonn.iz-soz.de

Zusammenfassung
Die Sonderfördermaßnahme CARMEN[1] ("Content Analysis, Retrieval and Metadata: Effective Networking") zielte unter anderem darauf ab, die Erweiterung von Recherchen in bibliographischen Fachdatenbanken ins Internet zu verbessern. Dabei war insbesondere die semantische Heterogenität zu behandeln, die durch unterschiedliche Inhaltserschließung in verschiedenen Datenbeständen auftritt. Dazu wurden verschiedene Ansätze wie Metadatenextraktion aus Internetquellen und Anfragetransfers über Cross-Konkordanzen und statistisch erzeugte Relationen gewählt. Dieser Aufsatz stellt die Konzepte und die Implementierung der Metadatenextraktion und der Anfragetransfers sowie die Evaluation der Auswirkungen auf das Retrievalergebnis vor.

Abstract
The project CARMEN ("Content Analysis, Retrieval and Metadata: Effective Networking") aimed among other goals at improving the expansion of searches in bibliographic databases into Internet searches. We pursued a set of different approaches to the treatment of semantic heterogeneity (meta-data extraction, query translation using statistic relations and cross-concordances). This paper describes the concepts and implementation of this approaches and the evaluation of the impact for the retrieval result.

[1] Gefördert durch das Bundesministerium für Bildung und Forschung im Rahmen des Programms „Global Info", FKZ 08SFC08 3.

1 Semantische Heterogenität

In der heutigen dezentralen Informationswelt werden Benutzer mit einer Vielzahl dezentraler Informationssysteme und Datenbestände mit ganz unterschiedlichen Inhaltserschließungsverfahren konfrontiert. In diesem Zusammenhang tritt semantische Heterogenität[2] auf, wenn unterschiedliche Datenbestände, die über eine Suchfunktion gemeinsam zugänglich gemacht werden, verschiedene Dokumentationssprachen benutzen, wenn Metadaten unterschiedlich oder überhaupt nicht erfasst werden oder wenn intellektuell aufgearbeitete Quellen mit in der Regel vollständig unerschlossenen Internetdokumenten zusammentreffen.

Standardisierungsbestrebungen wie die der Dublin Core Metadata Initiative sind eine wichtige Voraussetzung für eine Verbesserung der Verbindung von Datenbeständen, aber sie erfordern ein hierarchisches Modell der Kooperation, das von allen Beteiligten akzeptiert wird. Wegen der verschiedenen Interessen der unterschiedlichen Partner lässt sich ein solches Modell kaum verwirklichen. Vielmehr ist davon auszugehen, dass durch „anarchische Tendenzen“ die Unterschiede bei der Erstellung, Erschließung und Verbreitung von Dokumenten eher zunehmen werden. [Krause '96, Krause/Marx '00]

Im Projekt CARMEN wurde dieses Problem einerseits durch die automatische Extraktion von Metadaten aus Internetdokumenten und andererseits durch Systeme zur Transformation von Anfragen über Cross-Konkordanzen und statistisch erzeugte Relationen angegangen.

Zunächst wird in diesem Aufsatz die Extraktion von Meta-Daten aus Internetquellen beschrieben (Kap. 2). Daran anschließend wird das Vorgehen bei der Erstellung von Cross-Konkordanzen und statistischen Relationen dargestellt (Kap. 3). Ähnliche Verfahren wurden in den Projekten AIR/PHYS [Biebricher et al. '88] verwendet und werden auch für die „EuroSpider“[3] -Systeme für multilinguales Retrieval genutzt. [Braschler/Schäuble '98, '00]

Die Verfahren zum Transfer von Anfragen (Kap. 4) berücksichtigen, dass eine Veränderung der Daten in den abzufragenden Datenbanken oft nicht möglich ist. Anfrageerweiterung wurde im Zusammenhang mit „relevance feedback“ diskutiert. [Harman '88] Hier dient es der Übersetzung zwischen De-

[2] Der Begriff semantische Heterogenität ist hier anders zu verstehen als in der Diskussion um die technischen Probleme der Behandlung verschiedener DBMS mit unterschiedlichen Schemata. [Bright et al. '94, Hull '97].

[3] http://www.eurospider.ch/.

skriptoren oder Notationen verschiedenen Dokumentationssprachen während des Retrievals. In Retrievaltests (Kap. 5) wurde die Auswirkung der Transferverfahren auf das Rechercheergebnis untersucht.

2 Extraktion von Meta-Daten

2.1 Ansatz

Das Ziel bei der Extraktion von Meta-Daten ist die Anreicherung schlecht oder gar nicht inhaltlich erschlossener Quellen mit wichtigen Meta-Daten, wie z.B. Autor, Titel, Schlagwörter oder Abstract, schon während des Einsammelns von Dokumenten – während der unten beschriebene Transfer von Anfragen zum Suchzeitpunkt durchgeführt wird. Diese generierten Meta-Daten werden den Dokumenten hinzugefügt und stehen dann für das Retrieval zusammen mit sicheren Meta-Daten zur Verfügung, sie haben allerdings ein niedrigeres Gewicht.

Die entwickelten Heuristiken für die Extraktion von Meta-Daten hängen stark von Dateiformaten, vom Fachgebiet, von Eigenschaften einer Website und vom Layout der Seiten ab. Ein zuverlässiger und fach- wie institutionenunabhängiger Ansatz ist bisher nicht verfügbar. So lange sich die Konventionen zum Erstellen von HTML-Dokumenten so schnell ändern und ein „semantisches Web“ nicht vorliegt, können nur beschränkte und temporär gültige Lösungen entwickelt werden.

Relevante Internet-Dokumente für die Mathematik liegen meist im PostScript-Format vor. Die hier erschlossenen Dokumente (Preprints und Dissertationen) werden also in einem unstrukturierten Dateiformat gespeichert, enthalten allerdings hochwertige Meta-Daten, die durch Schlüsselworte und Layout-Informationen erkannt werden können. Durch die Anwendung von Werkzeugen wie „Prescript“[4] der New Zealand Digital Library wurde es möglich, die PostScript-Dokumente zu transformieren und zu analysieren. Dabei wurden der Abstract, Schlagworte und Klassifikationen erfolgreich extrahiert.

Internet-Dokumente aus den Sozialwissenschaften liegen in aller Regel als HTML-Dokumente vor, also in einem strukturierten Dateiformat. Allerdings werden die Möglichkeiten, mit HTML Inhalte auszuzeichnen, oft nur für die Formatierung genutzt. Meta-Tags werden kaum genutzt, und die Dokumente sind syntaktisch oft fehlerhaft. Verschiedene Institutionen benutzen unter-

[4] http://www.nzdl.org/html/prescript.html.

schiedliche Stile bei der Erstellung ihrer Dokumente, und viele Dokumente enthalten nicht einmal Informationen über den Autor oder die Institution selbst. Dadurch wird die Extraktion von Meta-Daten sehr erschwert.
Da die Analyse der (häufig fehlerhaften) HTML-Dokumente sehr fehleranfällig ist, wurden die Dokumente in XHTML konvertiert und dabei syntaktisch bereinigt. Dadurch wurde es möglich, bei der Extraktion bereits vorhandene XML-Werkzeuge zu nutzen und die Heuristiken mit der Anfragesprache XPath zu erstellen. [Strötgen/Kokkelink '01].

Das folgende Beispiel stellt den Algorithmus für die Extraktion und die Gewichtung des Titels dar (<x> gibt dabei eine interne Methoden-Nummer an, [x] das Gewicht der Meta-Daten). Zugrunde liegt dabei eine vorher durchgeführte Vorstrukturierung der Dokumente mit XPath.

```
If (<title> vorhanden && <title> enthält nicht „untitled“ && HMAX
    vorhanden){
    /* 'enthält nicht „untitled“' wird case insensitive im <titel>
       als Substring gesucht */
    If (<title>==HMAX) {
      <1> Titel[1,0]=<title>
    } elsif (<title> enthält HMAX) {
      /* 'enthält' meint hier immer case insensitive als Substring */
      <2> Titel[0,8]=<title>
    } elsif (HMAX enthält <title>) {
      <3> Titel[0,8]=HMAX
    } else {
      <4> Titel[0,8]=<title> + HMAX
    }
  } elsif (<title> vorhanden && <title> enthält nicht „untitled“ &&
     S vorhanden) {
     /* d.h. <title> vorhanden UND es existiert ein Eintrag S mit
      //p/b, //i/p usw. */
      <5> Titel[0,5]=<title> + S
  } elsif (<title> vorhanden) {
      <6> Titel[0,5]=<title>
  } elsif (<Hx> vorhanden) {
      <7> Titel[0,3]=HMAX
  } elsif (S vorhanden)
  {
      <8> Titel[0,1]= S
    }
  }
```

Abbildung 1: Heuristik für die Extraktion von Titeln

2.2 Evaluation

Für die Sozialwissenschaften wurde ein Testkorpus mit 3661 HTML-Dokumenten von den Web-Servern verschiedener relevanter Institutionen gesammelt. Von diesen Dokumenten enthielten 96% einen korrekt ausgezeich-

neten Titel; 17,7% der übrigen Dokumente enthielt einen fehlerhaft codierten Titel, die übrigen überhaupt keinen. Nur 25,5% enthalten Schlagworte, die alle korrekt ausgezeichnet wurden. Nur 21% enthalten einen richtig ausgezeichneten Abstract, 39,4% der übrigen Dokumente enthalten einen Abstract, der auf andere Weise im Dokument eingebaut wurde. Diese Voruntersuchung ist die Grundlage der Evaluation – denn da keine Methoden zur automatischen Inhaltserschließung angewendet werden sollten, können Meta-Daten nur da extrahiert werden, wo sie auch im Dokument vorhanden sind.

Für die Evaluation wurde eine repräsentative Stichprobe im Umfang von 360 Dokumenten gebildet, für die intellektuell die Relevanz der extrahierten Meta-Daten bestimmt wurde. Dabei wurden vier Relevanzstufen genutzt: hohe Präzision und Vollständigkeit, hohe Präzision für einen Teil der Extraktion, aber unvollständig und/oder auch Extraktion nicht relevanter Anteile; nicht relevant; nicht bewertbar.

Von den extrahierten Titeln sind 80% von mittlerer oder hoher Qualität, fast 100% der gefundenen Schlagworte sind von hoher Qualität und etwa 90% der extrahierten Abstracts sind von hoher oder mittlerer Qualität. [Binder et al. '02]

3 Semantische Relationen

Semantische Relationen zwischen Elementen von Dokumentationssprachen oder auch von solchen Elementen zu Freitexttermen wurden im Projekt CARMEN intellektuell erzeugt („Cross-Konkordanzen“) und über statistische Verfahren generiert.

3.1 Cross-Konkordanzen

Für die Bereiche Mathematik, Physik und Sozialwissenschaften wurden von Fachexperten intellektuell semantische Relationen zwischen verschiedenen Thesauri und Klassifikationen erstellt. Diese Verknüpfungen verbinden jeweils zwei Dokumentationssprachen miteinander, die Relationen wurden als Äquivalenz, Ober-/Unterbegriffsrelation und Ähnlichkeitsrelation erfasst und mit Gewichten (hoch, mittel, niedrig) versehen.

Für die Erstellung derartiger Cross-Konkordanzen wurden zwei Werkzeuge eingesetzt: das Web-basierte CarmenX[5] für die Konkordanzen zwischen Klas-

[5] http://www.bibliothek.uni-regensburg.de/projects/carmen12/.

sifikationen und das als semantisches Netzwerk organisierte SIS/TMS[6] für die Relationen zwischen Thesauri.

Einmal erarbeitete Konkordanzen ermöglichen sichere Übergänge zwischen verschiedenen Erschließungssystemen, ihre Erstellung und Wartung ist aber mit hohem finanziellen und zeitlichen Aufwand verbunden. Außerdem sind viele Dokumente, vor allem im Internet, überhaupt nicht mit einem kontrollierten Vokabular erschlossen. Daher sind ergänzend oder alternativ zusätzliche automatische Verfahren zu nutzen.

3.2 Statistisch erzeugte Relationen

Statistische Methoden ermöglichen die Erzeugung von semantischen Relationen auf der Grundlage vorhandener Dokumentbestände. Im beschriebenen Projekt wurde die Analyse von Wort-Kookkurrenzen gewählt. In den vergangenen 20 Jahren wurden vor allem im Kontext der automatischen Inhaltserschließung verschiedene Verfahren zur Kookkurrenzanalyse erprobt, [Biebricher et al. '88, Ferber '97] wobei die bedingte Wahrscheinlichkeit und der Äquivalenzindex mit die besten Ergebnisse lieferten und daher auch hier Anwendung fanden. Voraussetzung ist ein Parallelkorpus, der zwei Dokumentationssprachen verbindet [Mandl '99].

Für die Erzeugung semantischer Relationen, die für einen Transfer zwischen Fachdatenbanken und Internetquellen genutzt werden können, lässt sich das benötigte Parallelkorpus nur mit Mühe bereitstellen. Internetquellen, insbesondere aus den Sozialwissenschaften, sind in aller Regel überhaupt nicht inhaltlich erschlossen, schon gar nicht mit einem kontrollierten Vokabular. Da sich erst recht kaum Internetdokumente finden lassen, die mit mehreren Dokumentationssprachen erschlossen sind, ließ sich über Internetquellen kein geeignetes Parallelkorpus direkt erstellen.

Ein weiteres Ziel war die Verbindung von Dokumentationssprachen mit Freitexttermen, die im Gegensatz zu einem Thesaurus oder einer Klassifikation nicht kontrolliert sind und daher einer Eingrenzung und Vorbehandlung bedürfen.

Aus dem Bereich Sozialwissenschaften wurde ein Testkorpus mit etwa 6000 HTML-Dokumenten von wissenschaftlichen Einrichtungen im Internet zusammengestellt. Über diesen Bestand wurde ein Parallelkorpus simuliert, in-

[6] http://www.ics.forth.gr/proj/isst/Systems/sis-tms.html.

dem den Dokumenten einerseits über einfache Verfahren der automatischen Inhaltserschließung Deskriptoren aus dem Thesaurus Sozialwissenschaften und andererseits über den Volltextindexierer Fulcrum SearchServer Freitextterme zugeordnet wurden. Die Freitextterme wurden mit einem Porter-Stemmer[7] vorbehandelt und über Schwellenwerte wurde die Zahl der zugeordneten Freitextterme für ein Dokument begrenzt. Auf diese Weise wurde ein Korpus bereitgestellt, dessen Dokumente durch zwei verschiedene (Dokumentations-)Sprachen erschlossen waren und der nun wie ein Parallelkorpus behandelt wurde (siehe Abb. 2).

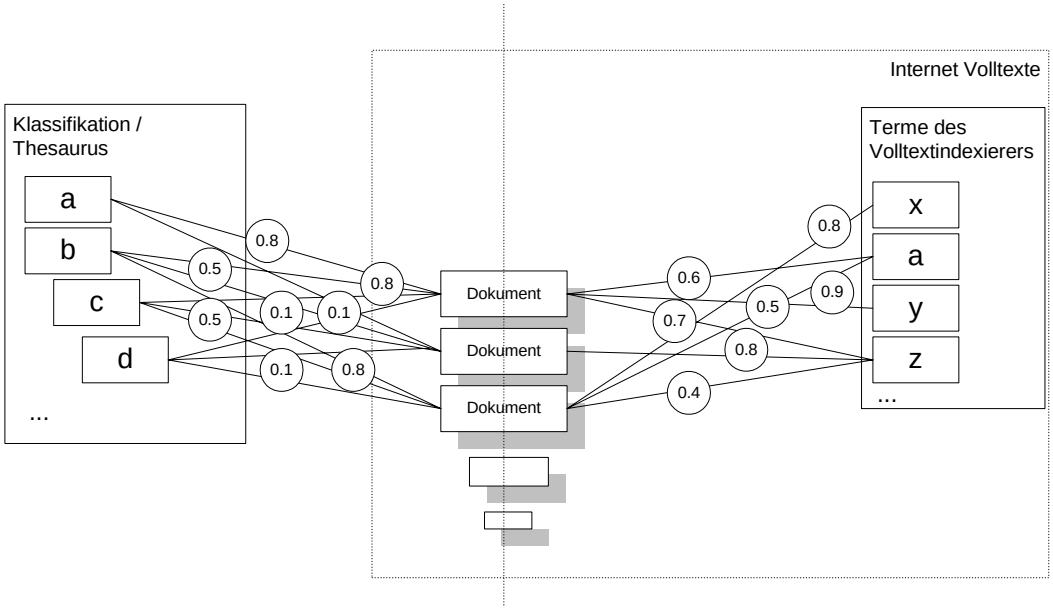

Abbildung 2: Parallelkorpus-Simulation mit vagen Deskriptoren und Volltexttermen

Für die Realisierung statistischer Transfermodule wurde auf das Werkzeug „Jester“ (**J**ava **E**nvironment for **S**tatistical **T**ransf**ER**s) zugegriffen, das im Kontext des Projekts Elvira II erstellt wurde. Jester führt die Berechnung der statistischen Transferbeziehungen durch und unterstützt dabei den Benutzer bei der Auswahl nötiger Parameter.

Als Maß für den statistischen Zusammenhang zweier Terme wird die „bedingte Wahrscheinlichkeit“, mit der zwei Terme gemeinsam auftreten, berechnet. Für einen Deskriptor **a** bestimmt **P(a)** die Wahrscheinlichkeit, dass **a** einem Dokument zugeordnet ist. Sie lässt sich ermitteln, indem die Zahl der Dokumente, in denen **a** auftritt durch die Gesamtzahl der Dokumente geteilt wird. **P(a ∧ b)** bezeichnet die Wahrscheinlichkeit, dass zwei Terme **a** und **b** gemeinsam auftreten. Ist das Auftreten der beiden Terme unabhängig vonein-

[7] Vgl. http://www.tartarus.org/~martin/PorterStemmer/.

ander, so sollte in etwa **P(a ∧ b) = P(a)* P(b)** gelten. Weichen diese Werte erheblich voneinander ab, so liegt ein systematischer Zusammenhang zwischen den Termen vor. Treten sie häufiger gemeinsam auf, als vorhergesagt, so sind sie wahrscheinlich nah verwandt. Treten sie deutlich seltener gemeinsam auf, so liegt nahe, dass sie eine gegensätzliche Beziehung haben und einander ausschließen.

Anstatt nur diese symmetrische Beziehung zu betrachten, die bei geeigneter Wahl eines Schwellenwertes die Ermittlung von Termpaaren erlaubt, die quasi synonym verwendet werden, kann auch die gerichtete bedingte Wahrscheinlichkeit betrachtet werden. Hier wird nur betrachtet, wie wahrscheinlich es ist, dass, wenn Deskriptor **a** vergeben wurde, ebenfalls Deskriptor **b** auftritt. **P(a) / P(a ∧ b)** beschreibt diesen Zusammenhang. Je größer dieser Wert ist, umso höher ist die gerichtete Abhängigkeit, die z.B. auftritt, falls **a** semantisch ein Unterbegriff von **b** ist.

Jester führt den Bearbeiter schrittweise durch die notwendigen Abläufe und unterstützt bei der Auswahl von Schwellenwerten. Auf diese Weise kann direkt bei der Manipulation der Parameter die Auswirkung beobachtet werden und interaktiv eine optimale Auswahl getroffen werden. Darüber hinaus stellt es eine Reihe von Heuristiken bereit, die bei der Behandlung systematischer Fehler wie Ausreißer angewendet werden können. [Hellweg '02]

Das Ergebnis der auf diesen Parallelkorpora durchgeführten Wort-Kookkurrenz-Analysen sind Term-Term-Matrizen, in der die gefundenen semantischen Relationen zwischen Deskriptoren und Freitexttermen für die spätere Nutzung beim Transfer von Anfragen bereitgehalten werden. [Hellweg et al. '01]

4 Transfer von Anfragen

Die semantischen Relationen waren nun für ein Suchsystem nutzbar zu machen. Da die genutzten verteilten Datenbanken nicht verändert werden sollten, sind die Relationen zur Manipulation der Suchanfragen zwischen Benutzerschnittstelle und Datenbank anzusiedeln. Die Suchanfrage soll also nicht - wie sonst in verteilten Sucharchitekturen, z.B. Meta-Suchmaschinen im WWW verbreitet - unverändert an die einzelnen Datenbanken weitergeleitet werden. Die Anfrage soll auch nicht einmalig erweitert und dann an alle Datenbanken gleich gestellt werden.

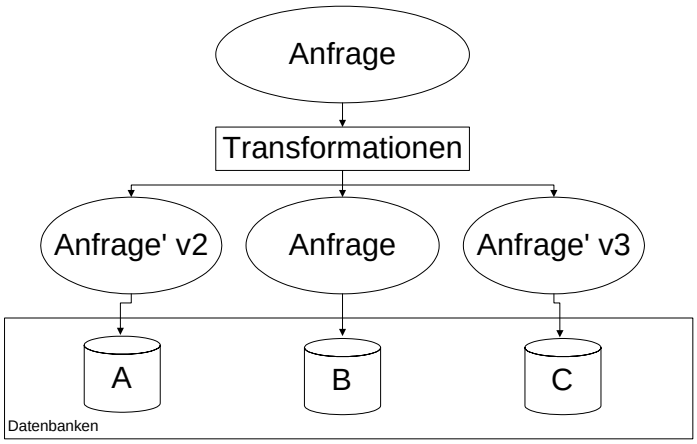

Abbildung 3: „Zwei-Schritt-Verfahren“

Stattdessen wird das im Informationszentrum Sozialwissenschaften entwickelte „Zwei-Schritt-Verfahren“[8] angewendet (siehe Abb. 3). Dabei werden datenbankspezifisch Anfragen generiert, die die jeweilige Inhaltserschließung der Datenbank berücksichtigen. Ist die Anfrage mit Hilfe einer Dokumentationssprache formuliert, die in einer der beteiligten Datenbanken genutzt wird, so wird die Anfrage unverändert an diese Datenbank gestellt. Für eine Datenbank, die eine andere Dokumentationssprache nutzt, wird die Anfrage über semantische Relationen zwischen beiden Dokumentationssprachen übersetzt.

Für das Projekt CARMEN waren die in Java entwickelten Softwaremodule, die diese Übersetzung der Anfrage leisten, in die Gesamtarchitektur einzubauen. Das an der Universität Dortmund entwickelte Retrievalsystem HyRex [Fuhr et al. '00] wurde dafür erweitert. Teilanfragen, für die ein Transfer angewendet werden kann (also z.B. nach Schlagwort) werden aus der in der Anfragesprache XIRQL vorliegenden Gesamtanfrage herausgetrennt. Die XIRQL-Teilanfrage wird in XML codiert und per Http-Request an den Servlet-Server gestellt, dort übersetzt und als Http-Response zurückgeschickt (siehe Abb. 3). In HyRex wird schließlich die veränderte Teilanfrage wieder in die Gesamtanfrage eingebaut. Sind mehrere Datenbanken beteiligt, ist dieser Schritt ggf. für jede Datenbank zu wiederholen.

[8] Angewendet in den Projekten ELVIRA, CARMEN, ViBSoz und ETB, vgl. http://www.gesis.org/Forschung/Informationstechnologie/Heterogenitaet.htm.

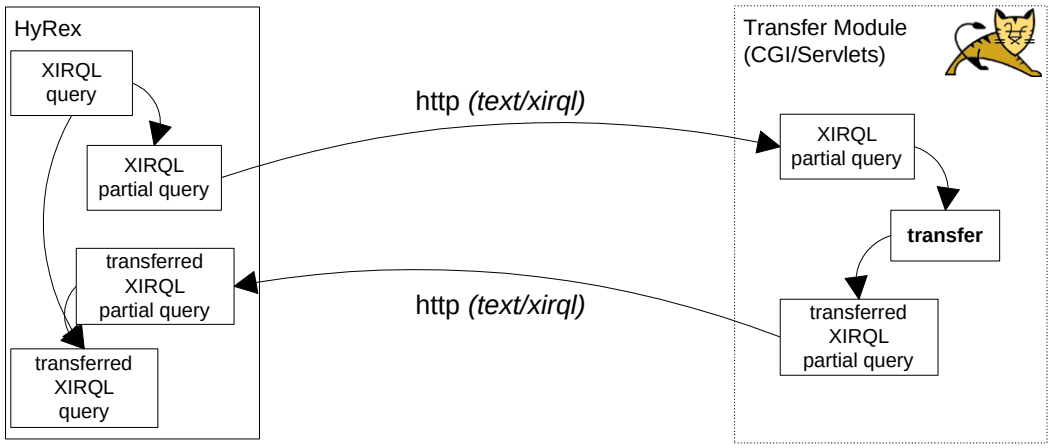

Abbildung 3: Anfrage-Transfer-Architektur

5 Evaluation

Für die Evaluation der Anfragentransfers über statistische Relationen wurden etwa 10000 HTML-Dokumente aus den Sozialwissenschaften mit dem Volltextindexierer Fulcrum indexiert (HyRex stand leider nicht rechtzeitig zum Test zur Verfügung). Die Ausnutzung der Gewichtsinformationen, die in HyRex genutzt werden, war dabei nicht möglich.

Das Test-Szenario geht von einer Suche in der Literaturdatenbank SOLIS[9] aus, die mit dem Thesaurus Sozialwissenschaften erschlossen ist. Eine Anfrage mit Deskriptoren aus diesem Thesaurus soll dann in einer Internet-Anfrage mit Freitexttermen erweitert werden. Die Thesaurus-Deskriptoren werden dabei also um Freitextterme erweitert.

Für den Test wurden zu drei Bereichen aus den Sozialwissenschaften (Frauenforschung, Migration und Industriesoziologie) jeweils zwei Anfragen gestellt. Dabei wurde zunächst mit der unveränderten SOLIS-Anfrage in Internet-Dokumenten gesucht und anschließend die transferierte Anfrage gestellt. Zwei Beispiele sollen hier kurz vorgestellt werden:

Eine Anfrage nutzte den Deskriptor „Dominanz” und lieferte 16 relevante Dokumente. Die transferierte Anfrage enthielt neu zusätzliche Terme: „Messen”, „Mongolei”, „Nichtregierungsorganisation”, „Flugzeug”, „Datenaustausch”, „Kommunikationsraum”, „Kommunikationstechnologie”, „Medienpädagogik”, „Wüste”. Die übersetzte Anfrage lieferte 14 zusätzliche Dokumente, von denen 7 relevant waren (50%, Zugewinn 44%). In diesem Beispiel

[9] http://www.gesis.org/Information/SOLIS/

konnten mit wenig Ballast zusätzliche relevante Dokumente gefunden werden.

Die genutzten Relationen sind mit großer Vorsicht zu interpretieren. „Wüste“ und „Mongolei“ werden nicht ergänzt, weil ein besonderes Dominanzproblem in der mongolischen Wüste angenommen wird. Es gibt lediglich ein Dokument, in dem eine Exkursion von Frauenrechtlerinnen nach China mit einem Zwischenstopp in der Mongolei beschrieben wird und dieses Dokument durch ungünstige Gewichtung einen hohen Einfluss auf die Relationen erhalten hat. Es gibt aber durchaus andere Fälle, in denen semantische Relationen durchaus Hinweise auf Problemfelder geben, z.B. bei statistischen Relationen zwischen mathematischen und physikalischen Klassifikationen, die mathematische Methoden mit physikalischen Anwendungsbereichen verbinden.

Ein anderes, weniger erfolgreiches Beispiel ist die Suche nach dem Deskriptor „Leiharbeit”. Zunächst wurden 10 relevante Dokumente gefunden. In der transferierten Anfrage wurden drei zusätzliche Terme hinzugefügt: „Arbeitsphysiologie”, „Organisationsmodell”, „Risikoabschätzung”. Von den 10 zusätzlich gefundenen Dokumenten waren nur 2 relevant (20%, Zugewinn 20%). In diesem Beispiel wurden mit erheblichem Ballast (80%) kaum zusätzliche Dokumente gefunden.

Zusammenfassend kann festgehalten werden, dass in allen transferierten Anfragen zusätzliche relevante Dokumente gefunden wurden. Die Precision der zusätzlichen Treffer liegt zwischen 13% und 55%. Ohne systematische Zusammenhänge bereits erkannt zu haben, wurden eher erfolgreiche und eher schwache Auswirkungen der Anfragetransfers vorgefunden.

6 Zusammenfassung und Ausblick

Es hat sich gezeigt, dass prinzipiell Meta-Daten extrahiert werden können. Die Extraktions-Werkzeuge wurden in den Gatherer des CARMEN-Projects „CARA“ eingebaut und stehen auch für andere Anwendungen zur Verfügung. Wegen der Kurzlebigkeit und des hohen Wartungsbedarfs der entwickelten Heuristiken scheint es aber fraglich, ob dieser hohe Aufwand aufgebracht werden kann.

Der Transfer von Anfragen unter Ausnutzung statistisch erzeugter Relationen hat sich grundsätzlich als brauchbar erwiesen, um die Ergebnisse der Suche zu verbessern. Allerdings bleiben einige Punkte offen. Zu klären ist beispielsweise, wie die Korpora und die Verfahren verbessert werden müssen,

um zu besseren Term-Term-Matrizen zu kommen. Außerdem wären die Anfragetransfers mit intellektuell erstellen Cross-Konkordanzen zum Vergleich heranzuziehen. Schließlich ist in echten Benutzertests zu evaluieren, welche Auswirkungen die Transfermodule im interaktiven Retrieval haben, wie sie von Benutzern sinnvoll parametrisiert werden können und welche Probleme und Irritationen beim Benutzer auftreten können.

Die im Projekt CARMEN erzeugten Softwaremodule und Term-Term-Matrizen werden interessierten Anwendern zur Verfügung gestellt. Im IZ Sozialwissenschaften finden sie in anderen Diensten wie ViBSoz und ETB Verwendung, weitere Dienste wie der Informationsverbund Bildung - Sozialwissenschaften - Psychologie werden folgen.

7 Literaturverzeichnis

[Biebricher et al. '88] Biebricher, P.; Fuhr, N.; Lustig, G.; Schwantner, M.; Knorz, G.: The Automatic Indexing System AIR/PHYS. From Research to Application. In: Chiaramella, Y. (Hrsg.): SIGIR'88, Proceedings of the 11th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Grenoble, France, June 13-15, 1988. ACM, 1988. S. 333-342.

[Binder et al. '02] Binder, G.; Marx, J.; Mutschke, P.; Strötgen, R.; Plümer, J.; Kokkelink, S.: Heterogenitätsbehandlung bei textueller Information verschiedener Datentypen und Inhaltserschließungsverfahren. (IZ-Arbeitsbericht; Nr. 24) Bonn: IZ Sozialwissenschaften, 2002.

[Braschler/Schäuble '98] Braschler, M.; Schäuble, P.: Multilingual Information Retrieval Based on Document Alignment Techniques. In: Nikolaou, C.; Stephanidis, C. (Hrsg.): Research and Advanced Technology for Digital Libraries, Second European Conference, ECDL '98, Heraklion, Crete, Greece, September 21-23, 1998, Proceedings. (Lecture Notes in Computer Science): Springer, 1998. S. 183-197.

[Braschler/Schäuble '00] Braschler, M.; Schäuble, P.: Using Corpus-Based Approaches in a System for Multilingual Information Retrieval. In: Information Retrieval 2000, 3, Nr. 3; S. 273-284.

[Bright et al. '94] Bright, M.W.; Hurson, A.R.; Pakzad, S.H.: Automated Resolution of Semantic Heterogeneity in Multidatabases. In: ACM Transactions on Database Systems (TODS) 1994, 19, Nr. 2; S. 212-253.

[Ferber '97] Ferber, R.: Automated Indexing with Thesaurus Descriptors: A Co-occurence Based Approach to Multilingual Retrieval. In: Peters, C.; Thanos, C. (Hrsg.): Research and Advanced Technology for Digital Libraries. First European Conference, ECDL '97, Pisa, Italy, 1-3 September, Proceedings. (Lecture Notes in Computer Science): Springer, 1997. S. 233-252.

[Fuhr et al. '00] Fuhr, N.; Großjohann, K.; Kokkelink, S.: CAP7: Searching and Browsing in Distributed Document Collections. In: Borbinha, J.L.; Baker, T. (Hrsg.): Research and Advanced Technology for Digital Libraries, 4th European Conference, ECDL 2000,

Lisbon, Portugal, September 18-20, 2000, Proceedings. (Lecture Notes in Computer Science): Springer, 2000. S. 364-367.

[Harman '88] Harman, D.: Towards Interactive Query Expansion. In: Chiaramella, Y. (Hrsg.): SIGIR'88, Proceedings of the 11th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Grenoble, France, June 13-15, 1988. ACM, 1988. S. 321-331.

[Hellweg '02] Hellweg, H.: Einsatz von statistisch erstellten Transferbeziehungen zur Anfrage-Transformation in ELVIRA. In: Krause, J.; Stempfhuber, M. (Hrsg.): Integriertes Retrieval in heterogenen Daten. Text-Fakten-Integration am Beispiel des Verbandinformationssystems ELVIRA. (Forschungsberichte des IZ Sozialwissenschaften) Bonn: IZ Sozialwissenschaften, 2002.

[Hellweg et al. '01] Hellweg, H.; Krause, J.; Mandl, T.; Marx, J.; Müller, M.N.O.; Mutschke, P.; Strötgen, R.: Treatment of Semantic Heterogeneity in Information Retrieval. (IZ-Arbeitsbericht; Nr. 23) Bonn: IZ Sozialwissenschaften, 2001.

[Hull '97] Hull, R.: Managing Semantic Heterogeneity in Databases. A Theoretical Perspective. In: ACM Symposium on Principles of Databases. Proceedings. ACM, 1997. S. 51-61.

[Krause '96] Krause, J.: Informationserschließung und -bereitstellung zwischen Deregulation, Kommerzialisierung und weltweiter Vernetzung ("Schalenmodell"). (IZ-Arbeitsbericht; Nr. 6) Bonn: IZ Sozialwissenschaften, 1996.

[Krause/Marx '00] Krause, J.; Marx, J.: Vocabulary Switching and Automatic Metadata Extraction or How to Get Useful Information from a Digital Library. In: Information Seeking, Searching and Querying in Digital Libraries: Pre-Proceedings of the First DELOS Network of Excellence Workshop. Zürich, Switzerland, December, 11-12, 2000. Zürich, 2000. S. 133-134.

[Mandl '99] Mandl, T.: Effiziente Implementierung von statistischen Assoziationen im Text-Retrieval. In: Ockenfeld, M.; Mantwill, G.J. (Hrsg.): Information und Region; 51. Jahrestagung der Deutschen Gesellschaft für Informationswissenschaft und Informationspraxis e.V., DGI, Hamburg, 21. bis 23. September 1999; Proceedings. (Tagungen der Deutschen Gesellschaft für Informationswissenschaft und Informationspraxis; 1) Frankfurt am Main: DGI, 1999. S. 159-172.

[Strötgen/Kokkelink '01] Strötgen, R.; Kokkelink, S.: Metadatenextraktion aus Internetquellen: Heterogenitätsbehandlung im Projekt CARMEN. In: Schmidt, R. (Hrsg.): Information Research & Content Management: Orientierung, Ordnung und Organisation im Wissensmarkt; 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; Proceedings. (Tagungen der Deutschen Gesellschaft für Informationswissenschaft und Informationspraxis; 4) Frankfurt am Main: DGI, 2001. S. 56-66.