Fachdatenbanken und Internet-Quellen: Rechercheüberstieg durch Anfragetransfer

Robert Strötgen

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

Abstract: Die Sonderfördermaßnahme CARMEN[1] zielte unter anderem darauf ab, die Erweiterung von Recherchen in bibliographischen Fachdatenbanken ins Internet zu verbessern. Dabei war das Problem der semantischen 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 das Konzept und die Implementierung der Anfragetransfers sowie die Evaluation der Auswirkungen auf das Retrievalergebnis vor.

1 Semantische Heterogenität

Zunehmend werden Benutzer mit einer Vielzahl dezentraler Informationssysteme und Datenbestände mit verschiedenen Inhaltserschließungsverfahren konfrontiert. Semantische Heterogenität[2] tritt hier auf, wenn über eine Suchfunktion integrierte Datenbestände 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 wegen der verschiedenen Interessen der unterschiedlichen Partner lässt sich ein von allen Beteiligten akzeptiertes hierarchisches Modell der Kooperation kaum verwirklichen. [KM00]

Im Projekt CARMEN wurde dieses Problem durch die automatische Extraktion von Metadaten aus Internetdokumenten und durch Systeme zur Transformation von Anfragen angegangen. Der vorliegende Aufsatz beschreibt den zweiten Ansatz.[3] Zunächst werden dabei die Verfahren zur Erstellung von Cross-Konkordanzen und statistischen Relationen,

[1] Gefördert durch das BMB+F im Rahmen des Programms „Global Info“, FKZ 08SFC08 3.

[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. [BHP94]

[3] Zur Metadatenextraktion siehe [SK01].

die bei der Anfragentransformation genutzt werden, beschrieben (Kap. 2). Ähnliche Verfahren wurden in den Projekten AIR/PHYS [BFL+ 88] verwendet und werden auch für die „EuroSpider“[4]-Systeme für multilinguales Retrieval eingesetzt. [BS00] Die Verfahren zum Transfer von Anfragen (Kap. 3) berücksichtigen, dass eine Veränderung der Daten in den abzufragenden Datenbanken oft nicht möglich ist. Anfrageerweiterung wurde im Zusammenhang mit „relevance feedback“diskutiert. [Har88] In diesem Kontext dient es der Übersetzung zwischen verschiedenen Dokumentationssprachen während des Retrievals. In Retrievaltests (Kap. 4) wurde die Auswirkung auf das Rechercheergebnis untersucht.

2 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.

2.1 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 gewichtet. Dafür wurden zwei Werkzeuge eingesetzt: das Webbasierte CarmenX[5] für Konkordanzen zwischen Klassifikationen und das als semantisches Netzwerk organisierte SIS/TMS[6] für Relationen zwischen Thesauri.

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

2.2 Statistisch erzeugte Relationen

Statistische Methoden ermöglichen die Erzeugung von semantischen Relationen auf der Grundlage vorhandener Dokumentbestände. Hier 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, [BFL+ 88, Fer97] wobei sich vor allem die bedingte Wahrscheinlichkeit und der Äquiva-

[4] http://www.eurospider.ch
[5] http://www.bibliothek.uni-regensburg.de/projects/carmen12/
[6] http://www.ics.forth.gr/proj/isst/Systems/sis-tms.html

lenzindex bewährten, die daher auch hier Anwendung fanden. Voraussetzung ist ein Parallelkorpus, der zwei (Dokumentations-)Sprachen verbindet. Für die Erzeugung semantischer Relationen, die Fachdatenbanken und Internetquellen verbinden können, lässt sich das benötigte Parallelkorpus nur mit Mühe bereitstellen. Internetquellen, insbesondere sozialwissenschaftliche, sind in aller Regel überhaupt nicht inhaltlich erschlossen, schon gar nicht mit mehreren kontrollierten Vokabularen. Ein Ziel war daher die Verbindung von Dokumentationssprachen mit (unkontrollierten) Freitexttermen.

Aus dem Bereich Sozialwissenschaften wurde ein Testkorpus mit etwa 6.000 HTML–Dokumenten zusammengestellt. Über diesen Bestand wurde ein Parallelkorpus simuliert. Dafür wurden den Dokumenten einerseits über einfache Verfahren der automatischen Inhaltserschließung Deskriptoren aus dem „Thesaurus Sozialwissenschaften“und andererseits über den Volltextindexierer Fulcrum SearchServer Freitextterme zugeordnet, die mit einem Porter-Stemmer vorbehandelt und deren Zahl über Schwellenwerte begrenzt wurde. 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. 1).

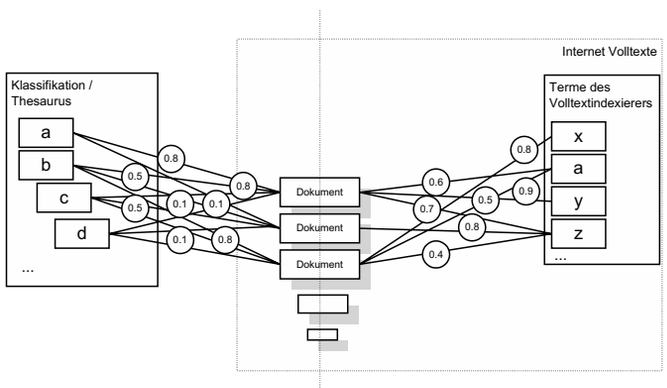

Abbildung 1: Parallelkorpus-Simulation mit vagen Deskriptoren und Volltexttermen

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.

3 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 Anfrage soll weder unverändert an die einzelnen Datenbanken weitergeleitet noch ein-

malig erweitert und dann an alle Systeme identisch gestellt werden. Stattdessen wird das so genannte „Zwei-Schritt-Verfahren“[7] angewendet. 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 Sprache nutzt, wird die Anfrage über semantische Relationen zwischen beiden Sprachen übersetzt.

Für das Projekt CARMEN wurden die in Java entwickelten Softwaremodule für Anfragetransfers in die Gesamtarchitektur eingebaut und dafür an das an der Universität Dortmund entwickelte Retrievalsystem HyRex[8] angebunden.

4 Evaluation

Für die Evaluation der Anfragentransfers über statistische Relationen wurden etwa 10.000 sozialwissenschaftliche HTML-Dokumente mit dem Volltextindexierer Fulcrum indexiert. Das Test-Szenario geht von einer Suche in der Literaturdatenbank SOLIS[9] aus. Die Anfrage mit Deskriptoren aus dem Thesaurus Sozialwissenschaften soll dann für eine Internet–Anfrage mit Freitexttermen erweitert werden.

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. Ein Beispiel soll hier kurz vorgestellt werden: Eine Anfrage nutzte den Deskriptor „Dominanz“und lieferte 16 relevante Dokumente. Die transferierte Anfrage enthielt 9 zusätzliche Terme[10] und lieferte 14 zusätzliche Dokumente, von denen 7 relevant waren (50%, Zugewinn 44%).

Die 6 Beispielanfragen 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.

5 Zusammenfassung und Ausblick

Der Transfer von Anfragen unter Ausnutzung statistisch erzeugter Relationen hat sich grundsätzlich als brauchbar erwiesen, um die Ergebnisse der Suche zu verbessern. Al-

[7] Am IZ entwickelt und angewendet in den Projekten ELVIRA, CARMEN, ViBSoz und ETB, vgl. `http://www.gesis.org/Forschung/Informationstechnologie/Heterogenitaet.htm`

[8] `http://ls6-www.informatik.uni-dortmund.de/ir/projects/hyrex/`

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

[10] „Messen“, „Mongolei“, „Nichtregierungsorganisation“, „Flugzeug“, „Datenaustausch“, „Kommunikationsraum“, „Kommunikationstechnologie“, „Medienpädagogik“, „Wüste“

lerdings 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 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 bei der Benutzung auftreten können.

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

Literaturverzeichnis

[BFL+ 88] Peter Biebricher, Norbert Fuhr, Gerhard Lustig, Michael Schwantner, and Gerhard Knorz. The Automatic Indexing System AIR/PHYS. From Research to Application. In Yves Chiaramella, editor, *SIGIR'88, Proceedings of the 11th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Grenoble, France, June 13-15, 1988*, pages 333–342. ACM, 1988.

[BHP94] M. W. Bright, A. R. Hurson, and Simin H. Pakzad. Automated Resolution of Semantic Heterogeneity in Multidatabases. *ACM Transactions on Database Systems (TODS)*, 19(2):212–253, 1994.

[BS00] Martin Braschler and Peter Schäuble. Using Corpus-Based Approaches in a System for Multilingual Information Retrieval. *Information Retrieval*, 3(3):273–284, 2000.

[Fer97] Reginald Ferber. Automated Indexing with Thesaurus Descriptors: A Co-occurence Based Approach to Multilingual Retrieval. In Carol Peters and Costantino Thanos, editors, *Research and Advanced Technology for Digital Libraries. First European Conference, ECDL '97, Pisa, Italy, 1-3 September, Proceedings*, volume 1324 of *Lecture Notes in Computer Science*, pages 233–252. Springer, 1997.

[Har88] Donna Harman. Towards Interactive Query Expansion. In Yves Chiaramella, editor, *SIGIR'88, Proceedings of the 11th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Grenoble, France, June 13-15, 1988*, pages 321–331. ACM, 1988.

[KM00] Jürgen Krause and Jutta Marx. 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*, pages 133–134. Zürich, 2000.

[SK01] Robert Strötgen and Stefan Kokkelink. Metadatenextraktion aus Internetquellen: Heterogenitätsbehandlung im Projekt CARMEN. In Ralph Schmidt, editor, *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, pages 56–66. DGI, Frankfurt am Main, 2001.