

Integration von Internetdatenbanken mit eingeschränkten Anfragemöglichkeiten *

Eike Schallehn Martin Endig

Fakultät Informatik, Otto-von-Guericke-Universität Magdeburg
PSF 4120, 39016 Magdeburg

{schallehn|endig}@iti.cs.uni-magdeburg.de

Zusammenfassung

Die Integration von Internetdatenbanken ist ein aktuelles Forschungsgebiet, welches auf neu entwickelten Lösungen aus dem Bereich der föderierten und mediatorbasierten Informationssysteme aufbaut. Eine neue Anforderung in diesem Bereich ergibt sich unter anderem aus den meist stark eingeschränkten Anfragemöglichkeiten dieser Quellen. Wir stellen in diesem Papier einen Ansatz vor, der als Erweiterung des Föderierungsdienstes FRAQL umgesetzt wurde. Mit dessen Hilfe können vorherrschende Anfragemöglichkeiten beschrieben und ausgewertet werden. Bei der Umsetzung wurde die Effizienz der Verteilung als eines der Hauptkriterien berücksichtigt.

1 Motivation

Die Integration von Internetdatenbanken dient der Verdichtung der großen Menge öffentlich zugänglicher Informationen und der Bereitstellung einheitlicher Zugriffsschnittstellen für Endnutzer oder andere Applikationen. Da diesbezüglich ein erheblicher Bedarf besteht, wurde dieses Gebiet in der letzten Zeit zu einem Forschungsschwerpunkt. Insbesondere kamen hierbei neue Entwicklungen aus dem Bereich der mediatorbasierten Systeme und föderierten Datenbanken zum Einsatz. Hierdurch konnten zahlreiche Probleme gelöst werden, andere aber blieben weitgehend ungelöst. Alle Ansätze haben eine mehrschichtige Architektur gemeinsam, in der ein Mediator oder Föderationsdienst die globale Zugriffsschnittstelle bereitstellt, Anfragen und deren Verarbeitung koordiniert und diese an Adapter oder Wrapper weiterleitet, welche die Schnittstelle für eine Quelle oder eine Klasse von Quellen bereitstellen.

Ein besonderes Problem stellt hierbei die Integration von Quellen mit unterschiedlichen und in keiner Form standardisierten Anfragemöglichkeiten dar, wie sie insbesondere bei Internetdatenbanken existieren. Generell läßt sich auf Grund der unterschiedlichen Datenhaltungsmechanismen, Datenmodelle und Zugriffsschnittstellen keine allgemeingültige Methode angeben, mit der die möglichen Anfrageausdrücke beliebiger Quellen beschrieben werden können.

Im Rahmen eines Projektes zur Integration digitaler Bibliotheken wurde hierzu der Föderationsdienst FRAQL [SCS00] um die entsprechende Funktionalität erweitert. FRAQL bietet eine objekt-relationale SQL Erweiterung sowie umfassende Mechanismen zur Konfliktlösung. Prinzipiell sind die im folgenden dargestellten Lösungen jedoch allgemein für gängige Mediatoren und Föderationsdienste anwendbar. In Abschnitt 3 werden am Beispiel bibliographischer Internetdatenbanken typische Anfragemöglichkeiten dargestellt, um dann in Abschnitt 4 generelle Anforderungen und eine mindestens zu unterstützende Funktionalität abzuleiten. Daraus folgend werden in Abschnitt 5 die verwendeten Konzepte vorgestellt.

*Die Forschung wurde im Rahmen des Projektes "Föderierungsdienst für heterogene Dokumentenquellen" (BMBF 08SFB031) gefördert.

2 Stand der Technik

Die gegenwärtig vielversprechendsten Ansätze zur Datenintegration basieren auf lose gekoppelten föderierten Datenbanken in Verbindung mit Multidatenbanksprachen und mediatorbasierten Systemen [DD99]. Anfragesprachen zur Unterstützung der Integration heterogener Datenquellen sind zum Beispiel Multidatenbanksprachen wie MSOL [GLRS93] und SchemaSQL [LSS96]. Beispiele für Systemimplementationen sind föderierte Datenbanken wie Pegasus [ASD⁺91] oder der IBM DataJoiner [VZ98] sowie mediatorbasierte Systeme wie TSIMMIS [GPQ⁺97]. Die Beschreibung von Anfragemöglichkeiten von Datenquellen ist zum Beispiel im Zusammenhang mit den Projekten TSIMMIS [LYV⁺98], Garlic [HKWY97] und Information Manifold [LRO96] diskutiert. Probleme mit dem Zugriff auf Internetdatenbanken mit SQL und anderen strukturierten Anfragesprachen, sowie weitere Aspekte wie die Extraktion von Daten aus semi-strukturierten Dokumenten ist unter anderem in [SH99], [Ade98] und [HL98] sowie zahlreichen Veröffentlichungen dargestellt.

3 Beispielszenario: Bibliographische Internetdatenbanken

Die Integration von bibliographischen Metadaten zu Literaturreferenzen im Rahmen heterogener Umgebungen ist ein relevantes Anwendungsszenario, in welchem die Beschreibung von Anfragemöglichkeiten entsprechender Informationsquellen eine entscheidende Rolle spielt. Gerade im WWW werden derzeit von einer Vielzahl von Anbietern entsprechende Metadaten über proprietäre Anfrageschnittstellen zur Verfügung gestellt. Um eine Integration der Daten aus dem WWW zu ermöglichen, ist es daher zunächst erforderlich, zu wissen, welche Anfragemöglichkeiten für die jeweiligen Metadaten durch die Quellen bereitgestellt werden. So erlauben einige Quellen beispielsweise nur eine separate Anfrage nach Autoren, Titel oder Schlüsselwörtern (Keyword). Andere Quellen stellen sehr komplexe Anfrageschnittstellen bereit, mit denen nicht nur eine Anfrage nach einzelnen Attributen ermöglicht wird, sondern auch komplexe Anfragen. Dabei ist eine „beliebige“ Kombination von Anfrageattributen mit Hilfe spezieller Operatoren möglich. In der Regel werden von einzelnen Anbietern auch unterschiedliche Anfrageschnittstellen bereitgestellt, um somit auf die Bedürfnisse verschiedener Benutzer einzugehen. Zwei bekannte Anbieter von bibliographischen Metadaten zu Literaturreferenzen sind der *LINK Information Service* vom Springer-Verlag [SV00] und die *Computer Science Bibliography* (DBLP) der Universität Trier [DBL00]. Von diesen Anbietern werden sowohl einfache als auch komplexe Anfrageschnittstellen zum Zugriff auf Metadaten zu Literaturreferenzen bereitgestellt. Während sich die Anfragemöglichkeiten bei DBLP aus-

Anbieter	Autor	Titel	Abstract	Keyword	Jahr	Seiten	Konferenz	Zeitung	Band	Number	Sprache
DBLP	✓	✓	–	–	✓	✓	✓	✓	✓	✓	–
Springer	✓	✓	✓	✓	–	–	–	✓	–	–	✓

Tabelle 1: Gegenüberstellung der Anfrageattribute von DBLP und Springer

schließlich auf einfache Metadaten beschränken, ist es bei Springer zusätzlich möglich, Anfragen auch auf Zusammenfassungen (Abstract) und Volltexten durchzuführen. Dabei sind innerhalb einer Anfrage verschiedene Kombinationen von Anfrageattributen, wie Autor, Titel, Jahr, usw., zulässig, wobei auch einzelne Einschränkungen zu beachten sind (vgl. Tabelle 1). So lassen sich beispielsweise bei DBLP nur eine begrenzte Anzahl an Autoren gleichzeitig oder bei Springer kein Erscheinungsjahr angeben. Wie

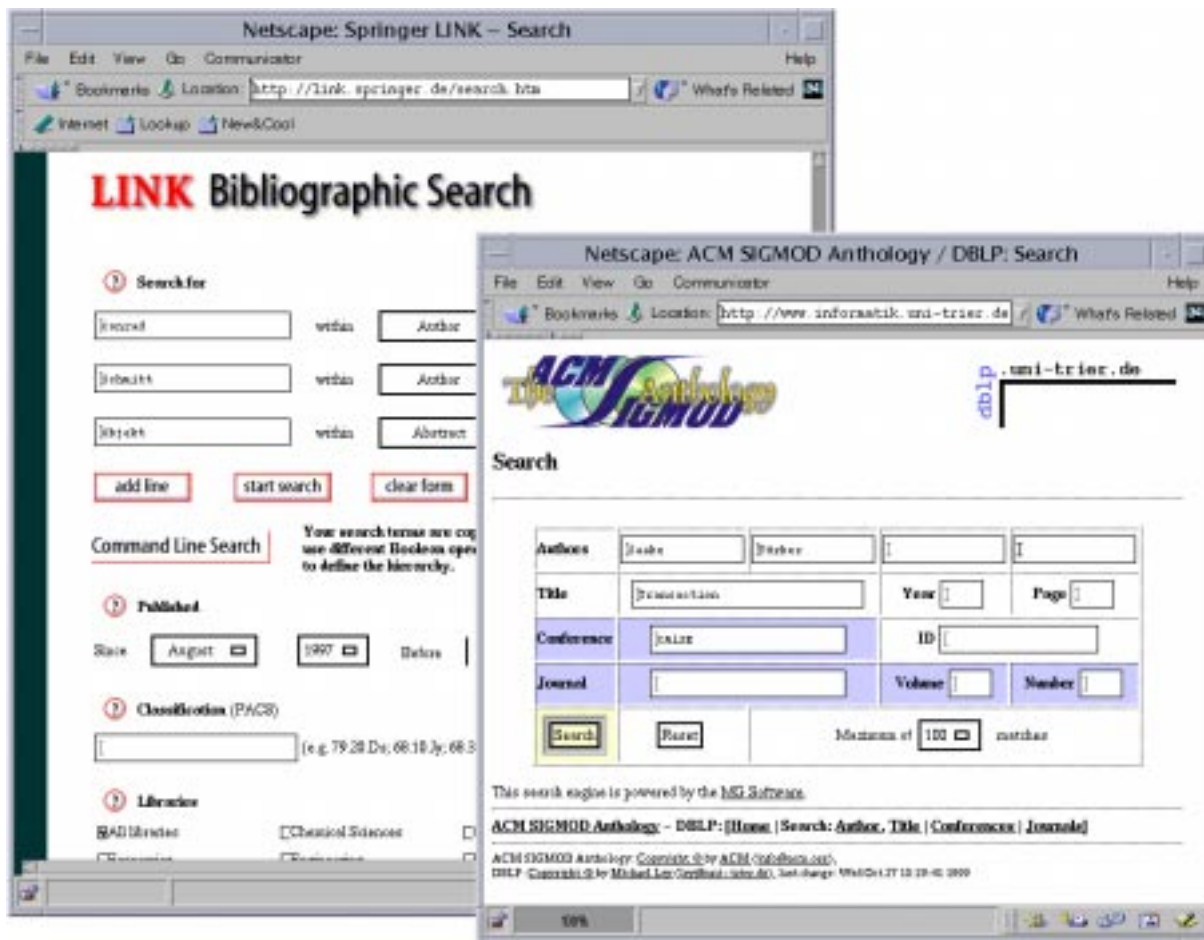


Abbildung 1: Beispiele für proprietäre Anfrageschnittstellen im WWW

diese einfachen Beispiele zeigen, ist es für die Integration von Metadaten in einem heterogenen Umfeld erforderlich, für Anfragen die Fähigkeiten der jeweiligen Quellen explizit zu beschreiben.

4 Anfragemöglichkeiten von Internetdatenbanken

Internetdatenbanken bieten heutzutage Anfrageschnittstellen auf der Basis strukturierter Anfragen an Datenbanksysteme oder zum Beispiel XML-Repositories, Information Retrieval-Systeme sowie häufig proprietärer Umsetzungen an. Deshalb ist es nicht möglich, eine allgemeine Methode zur Beschreibung von Anfragemöglichkeiten aller denkbaren Quellen anzugeben. Auf Grund des generellen Vorgehens bei Anfragen und gegebener Konventionen für die Nutzerunterstützung für die Suche kann man jedoch davon ausgehen, daß eine gewisse Kernfunktionalität bei allen Systemen gegeben ist.

Die grundsätzliche Zielstellung bei der Beschreibung von Anfragemöglichkeiten ist es, jene Anfragen zur Datenquelle weiterreichen zu können, die das zu übertragende Datenvolumen der Antwort minimieren. So wird einerseits der Kommunikationsaufwand vermindert, und andererseits die Weiterverarbeitung im Förderierungsdienst oder Mediator erleichtert. Hier kann oft keine umfassende Optimierung mehr erfolgen, da zum Beispiel Indexstrukturen fehlen.

Wir unterscheiden im folgenden grundsätzliche Klassen von Operationen, die von Internetdatenbanken nur zu einem gewissen Grad unterstützt werden.

Operationen zur Strukturierung von Ergebnissen, wie zum Beispiel die aus relationalen Bereich bekannten Projektionen, Gruppierungen, Duplikateliminierung usw. werden in der Regel nicht unterstützt. Die Anfrageergebnisse werden meist in einem statischen Format dargestellt, welches vom Nutzer nicht beeinflusst werden kann. Auch wenn diese Operationen zu einer Verminderung des Datenvolumens beitragen können, ist das Verhältnis in Standardanwendungsfällen lediglich ein konstanter Faktor.

Verbundoperationen werden meist ebenfalls nicht explizit angeboten. Sie werden meist durch einen internen Verbund und eine entsprechende Denormalisierung der Ergebnisse oder durch Navigationsmöglichkeiten angeboten. Hierbei liegt es am Adapter, die eventuell notwendigen Normalisierungen durchzuführen, und somit die weitere Anfragebearbeitung zu ermöglichen.

Selektionen werden von allen Internetdatenbanken unterstützt, jedoch meistens nur Konstantenselektionen, bei denen ein Attribut über einen Vergleichsoperator mit einem anzugebenden konstanten Wert verglichen werden kann. Generell können folgende Eigenschaften festgehalten werden:

- Quellen unterstützen lediglich eine eingeschränkte Menge von Prädikaten, d.h. sowohl die Menge der abfragbaren Attribute als auch die erlaubten Vergleichsoperatoren sind eingeschränkt.
- Prädikate können auf einfach Art und Weise logisch verknüpft werden. Selten besteht die Möglichkeit, komplexe, geschachtelte Bedingungen anzugeben. Am häufigsten werden einfache konjunktive Verknüpfungen unterstützt.
- Die Kombinierbarkeit der Prädikate ist begrenzt. So ist es oft der Fall, daß in einer Anfrage nur exklusiv ein Attribut verwendet werden kann. Weitere Beschränkungen bestehen bezüglich der Anzahl des Auftretens eines Attributs in einer Anfrage, da zum Beispiel die entsprechenden Anfragemasken nur eine begrenzte Anzahl von Feldern vorsehen.

Selektionen haben den größten Einfluß auf die physische Größe des Ergebnisses. Informationen und Heuristiken zur Selektivität von Prädikaten können für eine Optimierung innerhalb des Förderierungsdienstes herangezogen werden.

Auf Grund dieser Betrachtungen, wurde in unserem Projekt eine einfache Möglichkeit zur Spezifikation von Anfragemöglichkeiten umgesetzt, welche die Beschreibung abfragbarer Attribute, darauf anwendbarer Vergleichsoperatoren sowie der erlaubten Kombinationen bereitgestellt. Hierdurch ist eine einfache Möglichkeit gegeben, Anfragen mit hoher Selektivität zu formulieren, die gleichzeitig von einer großen Anzahl Quellen unterstützt wird.

5 Beschreibung und Auswertung von Anfragemöglichkeiten

Ausgehend von einem Förderierungsdienst, der eine objekt-relationale Datenbankschnittstelle anbietet, gehen wir derzeit von zwei Klassen von Quellen aus. Ist das Quellsystem ein relationales oder objekt-relationales Datenbanksystem, existieren keine Einschränkungen und die für die Quelle umgeschriebene Anfrage kann dort direkt bearbeitet werden. In der zweiten Klasse sind zum Beispiel Internetdatenbanken mit einfachen Web-Schnittstellen und den daraus resultierenden, oben eingeführten Beschränkungen. Sei \mathcal{R} die Menge der Namen von Relationen, welche von der Quelle S exportiert werden, und sei $Attr$ der abfragbaren Attribute der Relationen in \mathcal{R} . Dann kann die Beschreibung der Anfragemöglichkeiten $SrcDesc$ wie folgt definiert werden.

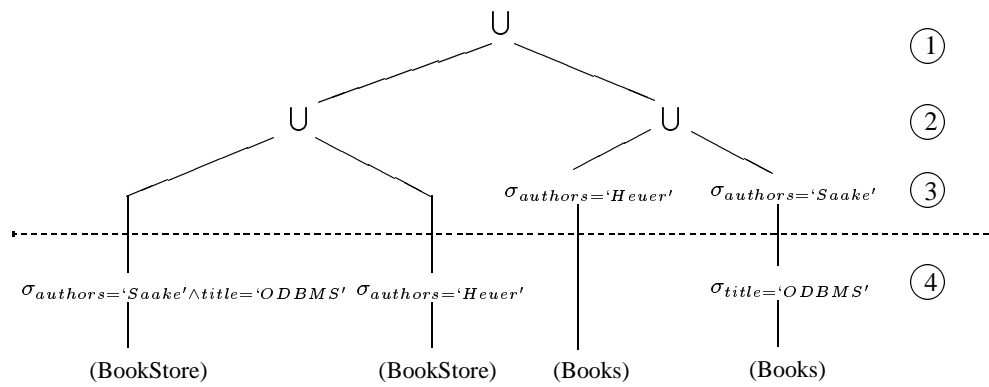


Abbildung 2: Auf der Basis der Quellenbeschreibung umgeschriebene Anfrage

$SrcDesc \triangleq \mathcal{R} \times RelDesc$	
$RelDesc \triangleq \mathbb{P}(Pred) \times ComDescr$	(Beschreibung der Importrelation R)
$Pred \triangleq Attr \times \mathbb{P}(\Theta)$	(unterstützte Konstantenselektionen)
$ComDescr \subseteq \mathbb{P}_{\mathcal{M}}(Attr)$	(mögliche Prädikatkombinationen)
$\Theta = \{>, \geq, =, \neq, <, \leq, LIKE\}$	(global unterstützte Vergleichsoperatoren)

Hierbei stellt $\mathbb{P}_{\mathcal{M}}(Attr)$ die Menge aller Multimengen dar, die aus der Menge der Attributnamen gebildet werden kann.

Als Beispiel betrachten wir die folgenden von zwei Quellen exportierten Relationen *Books* und *BookStore*, die der Einfachheit halber beide einem globalen Typ *Publications* entsprechen. Quelle 1 exportiert die Relation *BookStore* und unterstützt den Test auf Gleichheit für die Attribute *author*, *title* oder eine Kombination von beiden.

$$SrcDesc_1 = (\{(BooksStore, \{(authors, \{=\}), (title, \{=\})\}), \{\{authors\}, \{title\}, \{authors, title\}\}\})$$

Die Quellenbeschreibung für Quelle 1 kann in FRAQL wie folgt angegeben werden.

```
alter table BookStore set query constraints (
  predicates ( (authors,=), (title,=) ),
  combinations ( (authors), (title), (authors,title) )
);
```

Quelle 2 exportiert die Relation *Books* und unterstützt den Test auf Gleichheit für *title*.

$$SrcDesc_2 = (\{(Books, \{(title, \{=\})\}), \{\{title\}\}\})$$

Entsprechend diesen Beschreibungen können Anfragen zerlegt werden, und der von der Quelle bearbeitbare Teil dorthin weitergeleitet werden. Wird die folgende Anfrage an eine Relation *Publ* gerichtet, welche die Relationen *Books* und *Bookstore* integriert, erfolgt die Anfragebearbeitung wie in Abbildung 2 dargestellt.

```
select *
from Publ
where title = 'ODBMS' and
       authors = 'Saake' or
       authors = 'Heuer';
```

Die Quellen bearbeiten alle Konstantenselektionen unterstützter Prädikate und Prädikatkombinationen entsprechend der Quellenbeschreibung (4). In diesem Beispiel gehen wir davon aus, daß ein komplettes Abrufen der Relation *Books* möglich ist, ohne daß ein Prädikat spezifiziert wurde. Dies ist nicht immer trivial, kann aber meist bei der Adapterimplementierung realisiert werden. Nicht unterstützte

Prädikate, und solche die wegen fehlender Kombinationsmöglichkeiten nicht in die Quellenselektion aufgenommen werden konnten, werden im Förderierungsdienst bearbeitet. Dies beinhaltet Attributselektionen, die sich lediglich auf diese eine Importrelation beziehen. Das Selektionskriterium für jede einzelne Quelle ist in disjunktiver Normalform (2), d.h. komplexere Selektionen werden als mehrere Quellenfragen umgesetzt. Ergebnisse aus verschiedenen Importrelationen werden durch eine Vereinigung oder mit einer zusätzlichen Anwendung der Erkennung von identischen Objekten (same-Objekte) zusammen mit einer Tupelverschmelzung durchgeführt.

In diesem Beispiel war lediglich eine globale Relation beteiligt. Weitere Operationen wie Verbund, Projektion, relationsübergreifende Attributselektionen sowie Konstantenselektionen, welche Attribute betreffen, die von einer Abbildungsfunktion oder Tupelverschmelzung verändert werden, müssen anschließend an die hier dargestellten Schritte ausgeführt werden.

6 Zusammenfassung und Ausblick

In diesem Papier haben wir unseren Ansatz zur Beschreibung der Anfragemöglichkeiten von Internetdatenbanken dargestellt. Dieser beinhaltet die Spezifikation möglicher Prädikate sowie deren unterstützte konjunktiven Verknüpfungen. Hierdurch kann eine große Anzahl existierender Quellen unterstützt werden. Weiterhin ist durch diese Vorgehensweise eine hohe Selektivität von Anfragen gewährleistet, wodurch das zu übertragende Datenvolumen minimiert wird.

Da es nicht Ziel des Ansatzes war, alle möglichen Anfragemöglichkeiten von Internetdatenbanken abzubilden, wäre weiterhin zu untersuchen, welche Erweiterungen sinnvoll und mit vertretbarem Aufwand realisierbar sind. Insbesondere die Unterstützung komplexerer Prädikatverknüpfungen durch Internetdatenbanken ist oft gegeben, führt jedoch auf der Ebene des Förderierungsdienstes zu komplexen Umschreibungen des Anfragebaums, der die einfache Umsetzung als disjunktive Normalform nicht mehr erlauben würde. Ein weiterer Anknüpfungspunkt ist die Untersuchung von Optimierungsmöglichkeiten. Die Auswahl von Prädikaten nach ihrer Selektivität wurde hier bereits angedeutet. Hierzu könnten auch Informationen über die Datenverteilung herangezogen werden.

Literatur

- [Ade98] A. Adelberg. NoDoSE - A Tool for Semi-Automatically Extracting Structured and Semistructured Data from Text Documents. In L. Haas and A. Tiwary, editors, *SIGMOD'98, Proc. of the 1998 ACM SIGMOD Int. Conf. on Management of Data, June 1–4, 1998, Seattle, Washington, USA*, volume 25 of *ACM SIGMOD Record*, pages 283–294. ACM Press, June 1998.
- [ASD⁺91] R. Ahmed, P. De Smedt, W. Du, W. Kent, M.A. Ketabchi, W. Litwin, A. Rafii, and M.-C. Shan. The Pegasus Heterogeneous Multidatabase System. *IEEE Computer*, 24(12):19–27, December 1991.
- [DBL00] DBLP. <http://www.informatik.uni-trier.de/ley/db/index.html>, 2000.
- [DD99] R. Domenig and K. R. Dittrich. An Overview and Classification of Mediated Query Systems. *SIGMOD Record*, 28(3), 1999.
- [GLRS93] J. Grant, W. Litwin, N. Roussopoulos, and T. Sellis. Query Languages for Relational Multidatabases. *VLDB Journal*, 2(2):153–171, 1993.

- [GPQ⁺97] H. Garcia-Molina, Y. Papakonstantinou, D. Quass, A. Rajaraman, Y. Sagiv, J. D. Ullman, V. Vassalos, and J. Widom. The TSIMMIS Approach to Mediation: Data Models and Languages. *Journal of Intelligent Information Systems*, 8(2):117–132, March/April 1997.
- [HKWY97] L. M. Haas, D. Kossmann, E. L. Wimmers, and J. Yang. Optimizing Queries Across Diverse Data Sources. In *VLDB'97, Proceedings of 23rd International Conference on Very Large Data Bases*, pages 276–285, 1997.
- [HL98] R. Himmelröder and B. Ludäscher. Querying the Web with FLORID. In M. H. Scholl, H. Riedel, T. Grust, and D. Gluche, editors, *Kurzfassungen — 10. Workshop “Grundlagen von Datenbanken”, Konstanz (02.06.–05.06.98)*, number 63, pages 94–98. Universität Konstanz, Fachbereich Informatik, 1998.
- [LRO96] A. Levy, A. Rajaraman, and J. Ordille. Querying Heterogeneous Information Sources Using Source Descriptions. In T. M. Vijayaraman et al., editors, *Proceedings of the Twenty-second International Conference on Very Large Data Bases, September 3–6, 1996, Mumbai (Bombay), India*, pages 251–262, Los Altos, CA 94022, USA, 1996. Morgan Kaufmann Publishers.
- [LSS96] L.V.S. Lakshmanan, F. Sadri, and I.N. Subramanian. SchemaSQL - A Language for Interoperability in Relational Multi-Database Systems. In T. M. Vijayaraman, A.P. Buchmann, C. Mohan, and N.L. Sarda, editors, *VLDB'96, Proc. of 22th Int. Conf. on Very Large Data Bases, 1996, Mumbai (Bombay), India*, pages 239–250. Morgan Kaufmann, 1996.
- [LYV⁺98] C. Li, R. Yerneni, V. Vassalos, H. Garcia-Molina, Y. Papakonstantinou, J. Ullman, and M. Valiveti. Capability Based Mediation in TSIMMIS. *SIGMOD Record (ACM Special Interest Group on Management of Data)*, 27(2), 1998.
- [SCS00] K. Sattler, S. Conrad, and G. Saake. Adding Conflict Resolution Features to a Query Language for Database Federations. In M. Roantree, W. Hasselbring, and S. Conrad, editors, *Proc. 3rd Int. Workshop on Engineering Federated Information Systems, EFIS'00, Dublin, Ireland, June*, pages 41–52, Berlin, 2000. Akadem. Verlagsgesellschaft.
- [SH99] K. Sattler and M. Höding. Adapter Generation for Extraction and Querying Data from Web Sources. In *Proc. of 2nd ACM SIGMOD Workshop WebDB'99*, 1999. <http://www-rocq.inria.fr/~cluett/webdb99.html>.
- [SV00] Springer-Verlag. <http://link.springer.de/search.htm>, 2000.
- [VZ98] S. Venkataraman and T. Zhang. Heterogeneous Database Query Optimization in DB2 Universal DataJoiner. In A. Gupta, O. Shmueli, and J. Widom, editors, *VLDB'98, Proc. of 24rd Int. Conf. on Very Large Data Bases, 1998, New York City, New York, USA*, pages 685–689. Morgan Kaufmann, 1998.