Der Dissertationspreis der GI

Die Gesellschaft für Informatik e.V. (GI) vergibt jährlich einen Preis für eine hervorragende Dissertation im Bereich der Informatik. Hierzu zählen nicht nur Arbeiten, die einen Fortschritt für die Informatik bedeuten, sondern auch Arbeiten aus dem Bereich der Anwendungen in anderen Disziplinen und Arbeiten, die die Wechselwirkungen zwischen Informatik und Gesellschaft untersuchen.

Die Auswahl stützt sich auf die von den Universitäten und Hochschulen für diesen Preis vorgeschlagenen Dissertationen. Jede dieser Hochschulen kann jedes Jahr nur eine Dissertation vorschlagen. Somit können die zum Auswahlverfahren der GI vorgeschlagenen Kandidatinnen und Kandidaten bereits als „Preisträger" ihrer Hochschule angesehen werden.

Im Rahmen eines Auswahlkolloquiums werden die vorgeschlagenen Dissertationen dem Nominierungsausschuss vorgestellt. Dabei beeindrucken mich immer wieder das wissenschaftlich hohe Niveau der Arbeiten, die hervorragende Qualität der Vorträge und die lebhaften Diskussionen. Entsprechend schwer ist es, aus dieser Auswahl eine einzige Dissertation zu bestimmen, die durch den Preis besonders gewürdigt wird.

Ich freue mich, dass mit der Präsentation der Preisträger und weiterer vorgeschlagener Dissertationen in der Rubrik *Ausgezeichnete Dissertationen* der neu gestalteten **it** die Ungerechtigkeit, eine aus mehreren ebenbürtigen Dissertationen hervorzuheben, etwas ausgeglichen und gleichzeitig ein Beitrag zum Wissenstransfer geleistet wird. Verwiesen sei auch auf den Band *Ausgezeichnete Informatikdissertationen* in der GI-Edition *Lecture Notes in Informatics*, in dem jährlich alle nominierten Dissertationen vorgestellt werden (siehe http://www.gi-ev.de/LNI).

Prof. Dr. Dorothea Wagner,
Universität Konstanz
Vorsitzende des
Nominierungsausschusses

Qualitätsgesteuerte Anfragebearbeitung für Integrierte Informationssysteme

Felix Naumann, Institut für Informatik, Humboldt-Universität zu Berlin, GI-Preisträger 2000

Die jüngere Entwicklung fort von zentralisierten Datenbankmanagementsystemen hin zu Informationssystemen, die verteilte, autonome Datenquellen integrieren, führt zu einer Verschiebung der Forschungsrichtung von der traditionellen *Anfrageoptimierung* zum Bereich der *Anfrageplanung*. Letztere untersucht das Problem, Ausführungspläne über verteilte, heterogene, überlappende und autonome Datenquellen zu finden. Das wichtigste Unterscheidungsmerkmal der unterschiedlichen Ausführungspläne ist nicht mehr wie für Datenbanksysteme die Antwortzeit, sondern die Informationsqualität des Resultats.

Informationsquellen auf dem World Wide Web sind hervorragende Beispiele solcher autonomen Quellen. Mit diesem neuen Typus der Informationsquellen verschieben sich auch die qualitativen Erwartungen der Anwender:

Die Dissertation wurde 2002 in *Lecture Notes in Computer Sciences* (LNCS) 2261, Springer Verlag, Heidelberg, veröffentlicht. Die Gutachter des Promotionsverfahrens waren Prof. Johann-Christoph Freytag (Humboldt Universität zu Berlin), Prof. Myra Spiliopoulou (Handelshochschule Leipzig) und Prof. Gio Wiederhold (Stanford University).

- Nutzer erwarten korrekte Ergebnisse, aber akzeptieren Datensätze, deren Werte *nahe* den Anfragebedingungen sind.
- Nutzer erhoffen von einem Informationssystem ein vollständiges Ergebnis, aber akzeptieren unvollständige Ergebnisse, z. B. bei eingeschränkten Ressourcen.
- Nutzer erwarten ein lückenloses Ergebnis, d. h. es sollte alle Attribute der Anfrage enthalten, und kein Attribut sollte null-Werte enthalten. Jedoch akzeptieren Nutzer lückenhafte Ergebnisse, also Ergebnisse mit fehlenden Attributwerten – eine unvollständige Antwort mit einigen Lücken ist besser als keine Antwort.

Um Nutzern die Informationen des Webs zu erschließen, werden integrierte webbasierte Informationssysteme entwickelt. Solche Systeme bieten Nutzern eine einheitliche und integrierende Schnittstelle für viele Informationsquellen. Die jüngst entwickelte WebService-Spezifikation ist ein weiterer Schritt in die Richtung integrierter Informationsysteme. Die Erweiterung von WebServices um die hier genannten Methoden ist ein vielversprechender Ansatz, die Integration autonomer Daten zu verbessern.

In unserer Methode zur qualitätsgesteuerten Anfragebearbeitung bewerten wir jede Quelle gemäß einer Menge von Qualitätskriterien wie Vollständigkeit, Verständlichkeit oder Genauigkeit, und stellen Algorithmen vor, die schnell qualitativ gute oder sogar optimale Anfragepläne finden, das heißt Ausführungsstrategien, die Resultate höchster Gesamtqualität produzieren.

Ein einfaches Anwendungsszenario eines solchen Systems ist eine Metasuchmachine, die existierende Suchmaschinen als autonome Quellen integriert. Andere Beispiele sind Aktieninformationsdienste, integrierte Wetterdienste oder verteilte molekularbiologische Datenbanken. Die folgenden Abschnitte dienen gleichsam als Rezept, um derartige Systeme mit Qualitätsüberlegungen zu integrieren.

1 Beschreibung von Datenquellen und Nutzeranfragen

Um einer Nutzeranfrage geeignete Datenquellen zuzuordnen, muss man beide auf kompatible Weise beschreiben. Die Beschreibungen müssen flexibel genug sein, Heterogenitäten zu überbrücken und so möglichst viele unterschiedliche Quellen einbinden zu können.

Wir verwenden das Konzept der universellen Relation, um sowohl Quellen als auch Anfragen zu beschreiben. Wir stellen das *UR-Tableau* als Darstellung der universellen Relation mit den konstituierenden Relationen, dazu passenden Quellen und schließlich Anfragen vor. Dieses Tableau erlaubt es einem System, einer Nutzeranfrage solche Quellen zuzuordnen, die prinzipiell eine Antwort – gleich welcher Qualität – liefern können.

2 Informationsintegration

Autonome Datenquellen überlappen sich extensional, also in den Objekten, die sie repräsentieren, und intensional, also in den Daten (Attributen) über die einzelnen Objekte. Um eine Nutzeranfrage unter diesen Bedingungen sinnvoll zu beantworten, bedarf es zweier Zutaten: Ein integriertes System muss erstens die Überlappungen erkennen und zweitens Datenkonflikte in den sich überlappenden Daten lösen.

Methoden der Objektidentifikation und des *Data Cleansing* vermögen Daten über gleiche Objekte zu erkennen. Nach der Objektidentifikation stehen oft mehrere Repräsentationen desselben Objekts zur Verfügung. Existieren zwei verschiedene Werte für das gleiche Attribut, herrscht ein Datenkonflikt. Zur Lösung von Datenkonflikten dient eine allgemeine *Konfliktlösungsfunktion,* die für konkrete Anwendungen spezialisiert werden muss.

3 Erstellung von Anfrageplänen

Um Nutzeranfragen zu beantworten, müssen Anfragepläne gefunden werden. Darin wird festgehalten, welche Quellen verwendet werden, um die Relationen mit Daten zu füllen.

Herkömmliche Forschungsansätze sind ungeeignet, da sie *alle* Pläne suchen, die die Anfrage prinzipiell beantworten können, auch wenn die Qualität der Resultate einzelner Pläne schlecht ist und viele Pläne redundant sind.

Wir stellen ein neues Paradigma der Anfragebearbeitung auf, welches diesen Mangel behebt. Drei neue Integrationsoperatoren erlauben die zielgerichtete Suche nach nur noch einem einzigen, qualitativ optimalen Plan. Die Vereinigung der Ergebnisse der einzelnen Pläne im herkömmlichen Paradigma wird nun bereits im Plan selbst mit Hilfe so genannter *Outerjoin*-Operatoren modelliert.

4 Definition von Informationsqualität

Wir interpretieren Konzepte wie „beste“ Pläne und „befriedigende“ Antworten als Pläne und Antworten von hoher Informationsqualität. Um diese zu bestimmen und zu maximieren, bedarf es eines Qualitätsmaßes.

Informationsqualität wird durch eine Menge einzelner Qualitätskriterien, etwa Vollständigkeit, Glaubwürdigkeit usw., definiert. Natürlich sind nicht alle Kriterien in jeder Situation relevant. Vielmehr findet für jede Anwendung eine Auswahl der Kriterien statt. Faktoren, die bei der Auswahl eine Rolle spielen, sind die Anwendungsdomäne, die Präferenzen voraussichtlicher Nutzer, der Anbieter des integrierten Informationssystems und schließlich die Fähigkeit, die Kriterien zu quantifizieren.

5 Bewertung von Datenquellen nach Qualitätskriterien

Um Qualitätskriterien für Anfrageplanung und Informationsintegration nutzen zu können, müssen sie quantifiziert werden. Wir identifizieren drei Quellen solcher Quantifikation:

- Nutzer können wertvolle Hinweise zur Informationsqualität geben, z. B. zur Verständlichkeit oder Reputation.
- Die Quellen selbst veröffentlichen – direkt oder indirekt – Qualitätsbewertungen wie Preis oder Aktualität.
- Der Anfrageprozess liefert Qualitätswerte wie Anfragezeit und Ergebnisumfang.

Das Ziel der Qualitätsbewertung ist ein Qualitätsvektor für jede Informationsquelle. Dieser Vektor enthält für jedes Kriterium eine Bewertung und dient als Grundlage für die Bewertung von Anfrageplänen.

6 Aggregation von Qualitätswerten

Um Nutzeranfragen an integrierte Systeme zu beantworten, ist es nötig, Daten von mehr als einer Quelle in einem Plan zu kombinieren. Um die Qualität der kombinierten Antwort (also auch die Qualität des Plans) zu bestimmen, muss man die Qualitätswerte der im Plan beteiligten Quellen aggregieren.

Individuelle *Merge-Funktionen* für Qualitätskriterien fassen die Qualitätswerte eines Kriteriums zu einem neuen Wert zusammen. Auf diese Weise werden Qualitätswerte entlang des Anfrageplans aggregiert, um so Qualitätswerte für den Gesamtplan zu erhalten. Mit einer solchen Strategie wird in herkömmlichen Datenbanksystemen die Anfragezeit eines Plans vorhergesagt.

7 Erstellung einer Rangordnung gemäß mehrerer Kriterien

Eine Datenquelle oder ein Anfrageplan wird mittels mehrerer Kriterien bewertet. Im Allgemeinen haben die Qualitätsmaße der Kriterien unterschiedliche Einheiten und Wertebereiche, außerdem sollen Nutzer die einzelnen Kriterien gewichten können. Um zu entscheiden, welche Quelle die beste oder welcher Plan der beste ist, müssen die einzelnen Werte zu einem Gesamtqualitätswert zusammengefasst werden. Die Literatur bietet viele solcher Methoden, von einfachen wie das *Simple Additive Weighting* bis hin zu komplexen wie die *Data Envelopment Analysis*.

8 Effiziente und effektive Anfrageplanung

Algorithmen zur Anfrageplanung sollen effizient zu einer Nutzeranfrage einen oder mehrere Anfragepläne finden. Die Dissertation stellt zwei Ansätze vor, die die besten *N* Pläne zu einer Anfrage finden. Der erste Ansatz ist eine Erweiterung gängiger Anfrageplanungsalgorithmen um eine nachgeschaltete qualitative Auswahl von Quellen und Plänen. Der zweite Ansatz integriert die Qualitätsbewertung mit der Anfrageplanung in einem Branch & Bound-Algorithmus – schon während der Planung werden wenig versprechende Teilpläne verworfen.

Die Notwendigkeit und Wichtigkeit von Qualitätsüberlegungen sind in der Forschung hinlänglich bekannt, werden jedoch oft ignoriert; die Integration dieser beiden Bereiche wurde bislang noch nicht vorgenommen. Wir behandeln umfassend alle Aspekte dieser Integration: Beginnend mit einem allgemeinen Modell der Anfrageplanung über eine Definition von Informationsqualität und einem Modell, Werte für Qualitätskriterien zu bestimmen, über Vergleichsmethoden, um die Qualitätsmaße auszuwerten, bis hin zu Algorithmen, die mit Hilfe dieser Methoden und Maße qualitativ hochwertige Ergebnisse auf Anfragen produzieren.

Daten, die aus autonomen Quellen – insbesondere aus WWW-Quellen – integriert werden, sind von schlechter Qualität. Diskussionen mit Wissenschaftlern und Praktikern vieler Gebiete haben gezeigt, dass niedrige Informationsqualität eines der dringendsten Probleme moderner Informationssysteme ist. Mit mehr und mehr verfügbaren Daten und mehr und mehr autonomen Quellen wird sich das Problem in Zukunft noch verschärfen. Wir hoffen, dass die Ergebnisse und Methoden dieser Arbeit Eingang in integrierte Informationssysteme finden, um so wieder die Fähigkeit zu erlangen, Nutzern auf effiziente Weise qualitativ hochwertige Daten liefern zu können. Diese Fähigkeit wurde einst verloren: beim Übergang von zentralisierten Datenbanken zu integrierten Informationssystemen.

Dr. rer. nat. Felix Naumann studierte von 1990 an Wirtschaftsmathematik an der Technischen Universität Berlin und schloss 1997 das Studium mit einem Diplom ab. Als Mitglied des Berlin-Brandenburger Graduiertenkollegs *Verteilte Informationssysteme* forschte Naumann von 1997 bis 2000 am Lehrstuhl von Prof. Johann-Christoph Freytag an der Humboldt Universität zu Berlin, und promovierte im Jahr 2000. Für seine Dissertation erhielt Naumann den Dissertationspreis 2000 der GI. 2001 und 2002 war er als Forscher am IBM Almaden Research Center in San Jose, USA beschäftigt. Ab 2003 kehrt er als Juniorprofessor für Informationsintegration an die Humboldt Universität zurück. Adresse: Humboldt Universität zu Berlin, Institut für Informatik, Unter den Linden 6, D-10099 Berlin, E-Mail: felix@hiqiq.de