Jan Hegewald

Informationsintegration in Biodatenbanken

VIEWEG+TEUBNER RESEARCH

Ausgezeichnete Arbeiten zur Informationsqualität

Herausgeber:

Dr. Marcus Gebauer

Bewertungskomission des
Information Quality Best Master Degree Award 2008:

Prof. Dr. Holger Hinrichs, FH Lübeck (Kommissionsvorsitz)
Dr. Marcus Gebauer, WestLB AG und Vorsitzender der DGIQ
Prof. Dr. Knut Hildebrand, HS Darmstadt
Bernhard Kurpicz, OrgaTech GmbH
Prof. Dr. Jens Lüssem, FH Kiel
Michael Mielke, Deutsche Bahn AG und Präsident der DGIQ
Prof. Dr. Felix Naumann, HPI, Uni Potsdam
Prof. Dr. Ines Rossak, FH Erfurt

Die Deutsche Gesellschaft für Informations- und Datenqualität e.V. (DGIQ) fördert und unterstützt alle Aktivitäten zur Verbesserung der Informationsqualität in Gesellschaft, Wirtschaft, Wissenschaft und Verwaltung. Zu diesem Zweck befasst sie sich mit den Voraussetzungen und Folgen der Daten- und Informationsqualität. Sie fördert zudem durch Innovation und Ausbildung die Wettbewerbsfähigkeit der Unternehmen sowie die des unternehmerischen und akademischen Nachwuchses in Deutschland.

Die vorliegende Schriftenreihe präsentiert ausgezeichnete studentische Abschlussarbeiten in der Daten- und Informationsqualität. Veröffentlicht werden hierin die Siegerarbeiten des jährlich stattfindenden „Information Quality Best Master Degree Award“.

Jan Hegewald

Informationsintegration in Biodatenbanken

Automatisches Finden von Abhängigkeiten zwischen Datenquellen

Mit einem Geleitwort von Dr. Marcus Gebauer

VIEWEG+TEUBNER RESEARCH

Bibliografische Information der Deutschen Nationalbibliothek
Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über <http://dnb.d-nb.de> abrufbar.

Gedruckt mit freundlicher Unterstützung
der Information Quality Management Group

1. Auflage 2009

Lektorat: Christel A. Roß

Vieweg+Teubner ist Teil der Fachverlagsgruppe Springer Science+Business Media.
www.viewegteubner.de

Umschlaggestaltung: KünkelLopka Medienentwicklung, Heidelberg
Druck und buchbinderische Verarbeitung: STRAUSS GMBH, Mörlenbach
Gedruckt auf säurefreiem und chlorfrei gebleichtem Papier.
Printed in Germany

ISBN 978-3-8348-0731-1

Geleitwort

Als Vorsitzender der Deutschen Gesellschaft für Informations- und Datenqualität (DGIQ e. V.) bin ich glücklich darüber, dass Sie dieses Buch in Ihren Händen halten. Das vorliegende Buch ist Ausdruck unseres Bestrebens, dem wissenschaftlichen Nachwuchs die Möglichkeit zu eröffnen, seine Arbeiten einem breiten Publikum darzustellen. Dass Sie gerade diese Arbeit vorfinden, ist Ergebnis eines strengen Auswahlprozesses, den die DGIQ mit dem zum ersten Mal ausgeschriebenen „Information Quality Best Master Degree Award 2008“ durchgeführt hat. Studenten waren aufgefordert, ihre Abschlussarbeiten zum Thema Informationsqualität in diesem Wettbewerb durch ihre begutachtenden Professoren einreichen zu lassen. Vertreter aus Wissenschaft, Forschung und Industrie haben diese akademischen Abschlussarbeiten begutachtet.

Jan Hegewald hat sich mit der vorliegenden Arbeit vorgenommen, Licht in den Dschungel verteilter Datenbestände zu bringen. An vielen Stellen der Welt, wie zum Beispiel im Human-Genom-Projekt, finden wir Informationen und Daten in den unterschiedlichsten Datenbanken und Datenbanksystemen. Eine Gesamtsicht auf solch verteilte Daten zu erhalten, ist in der Regel nur mit manuellem Aufwand und der menschlichen Intuition und Interpretation möglich. Dies ist allerdings häufig inakzeptabel langsam und aufwändig. Mit seiner Arbeit stellt der Preisträger ein „fast automatisches Verfahren“ vor, um identische realweltliche Objekte in verschiedenen Datenquellen effizient zu erkennen. Dies ist ihm bewundernswert originell, auf einem mathematisch festen Fundament gelungen.

Besonders freue ich mich, dass wir mit dem Verlag Vieweg+Teubner nun die Siegerarbeiten in einer Schriftenreihe jährlich veröffentlichen können. Für die Initiative des Verlages möchte ich mich recht herzlich bedanken.

Offenbach, 27. August 2008

Dr. Marcus Gebauer

Vorwort

Die moderne Informationstechnik ermöglicht es uns Daten auf allen Gebieten und in fast unbegrenzten Mengen zu sammeln. Doch wie schon die Neuronen in unserem Gehirn vor allem auf Grund Ihrer hochgradigen Vernetzung so etwas komplexes wie das Denken ermöglichen, lassen sich auch durch die Verknüpfung von digital gesammeltem Wissen ganz neue und weitergehende Erkenntnisse gewinnen. Dieses Buch leistet einen kleinen Beitrag zur Integration von Daten aus verschiedenen Datenquellen. Der konkrete Anwendungsfall ist die Molekularbiologie. Der vorgestellte Algorithmus ist jedoch auch in ganz anderen Bereichen anwendbar, wo es darum geht, gleiche Objekte in verschiedenen Datenbanken zu identifizieren. Erst vernetztes Wissen schafft einen größeren Kontext und ermöglicht es, über den Tellerrand zu schauen. Derart vernetztes Wissen kann dazu beitragen die wissenschaftliche Forschung entscheidend voranzubringen – beispielsweise beim Erforschen und Bekämpfen von Krankheiten.

Freilich ist trotz – oder gerade wegen – all der Möglichkeiten auch Wachsamkeit angebracht. Informationen, die sich über einen Menschen beispielsweise im Web finden lassen, führen – gekonnt verknüpft – schnell zu ausführlichen Profilen und einem umfassenderen Bild als derjenige es vielleicht gerne hätte. Die Privatsphäre eines Kunden mutiert durch einfach auszuwertende Konsumdaten, ergänzt um andere Verhaltensinformationen, schnell zur öffentlichen Sphäre.

Ich hoffe also, dass die Informationsintegration zu Fortschritten in der Wissenschaft beiträgt, die vor allem der Allgemeinheit zu Gute kommen.

Dieses Buch ist aus meiner Diplomarbeit entstanden. Einigen Personen, die es so weit haben kommen lassen, gebührt Dank.

Zunächst möchte ich mich bei Prof. Dr. Felix Naumann vom Hasso-Plattner-Institut in Potsdam und bei Prof. Dr. Ulf Leser von der Humboldt-Universität zu Berlin dafür bedanken, dass sie mir in einer bis dahin für mich etwas ungünstigen Situation die Möglichkeit und anschließend die Unterstützung zu dieser Arbeit gaben. Außerdem hat Felix Naumann mich während meines gesamten Studiums gefördert und er hatte auch die Idee meine Arbeit bei der DGIQ einzureichen. Jana Bauckmann – meiner Diplomarbeits-Betreuerin und Urheberin eines Algorithmus, den ich als Ausgangspunkt nahm – möchte ich für die vielen Anregungen und die Arbeit, die ich ihr gemacht habe, danken. Durch nächtelanges Korrekturlesen und Verbessern von manchmal unverständlichen Satzkonstruktionen, die

eine wahrscheinlich noch unverständlichere Materie zum Gegenstand hatten, hat Ricarda König dazu beigetragen der Arbeit einen letzten Schliff zu geben.

Eine unerwartete Freude war es für mich, als ich erfuhr, dass die DGIQ meine Arbeit mit dem 1. Platz des IQ Best Master Degree Award ausgezeichnet hat. Dass als Folge davon sogar einmal dieses Buch erscheinen würde, hätte ich damals nicht im Traum vermutet. Der DGIQ und hier besonders Dr. Marcus Gebauer danke ich deshalb ganz herzlich für das Vertrauen und die Anerkennung, die sie mir entgegen bringen!

Die Erstellung dieses Buches aus meiner Diplomarbeit wären ohne die sehr kreative und biowissenschaftlich fundierte Unterstützung von Maria Trusch nicht annähernd so gut gelungen. Als angehende Doktorin der Biochemie versteht sie im Gegensatz zu mir sogar, was all die Informationen in den molekularbiologischen Datenbanken genau bedeuten. Lieben Dank für die Hilfe!

Schlussendlich danke ich Ihnen, lieber Leser, für Ihr Interesse an der Datenintegration – viel Spaß beim Lesen!

Berlin, Oktober 2008

Jan Hegewald

Inhaltsverzeichnis

1 Einleitung

Die Biowissenschaften, auch als Life Sciences bezeichnet, haben in den letzten Jahren große Fortschritte gemacht: die Entschlüsselung des menschlichen Genoms, die Überwachung von Seuchen oder die systematische Erforschung von Krankheitsursachen sind nur einige Beispiele. Alle drei haben gemeinsam, dass sie zum Teil erst durch den Einsatz von IT-Systemen möglich wurden. Was Informationssysteme hierbei vor allem leisten, ist das Speichern und Analysieren großer Datenbestände.

Es existieren eine Reihe von Datenbanken, die Erkenntnisse einzelner Forschungsgebiete enthalten. Ein Beispiel ist etwa die *Protein Data Bank* (PDB)[1], die Proteine und deren Eigenschaften erfasst. *CATH*[2] ist eine Datenbank, die Proteine anhand ihrer Struktur hierarchisch klassifiziert. *SCOP*[3] hat einen ähnlichen Zweck.

Die eigenständigen Datenbanken lassen sich meist gut durchsuchen. Woran es teilweise mangelt, ist eine einheitliche Gesamtsicht auf thematisch verwandte Daten. Oft ist es erforderlich aus einem bestimmten Kontext auf Daten einer anderen Datenbank zuzugreifen. Momentan muss dies manuell erfolgen, beispielsweise indem die Bezeichnung eines in einer Datenbank enthaltenen Proteins notiert wird und anschließend anhand der Bezeichnung nach entsprechenden Informationen in einer anderen Datenbank gesucht wird. Eine integrierte, datenbankübergreifende Sicht auf die Daten existiert nicht, würde die Effizienz der Forschungsarbeit aber enorm erhöhen. Skalierbare Integrationsarchitekturen werden daher dringend benötigt um die stetig wachsenden Datenmengen analysieren zu können [Sin05].

Hier setzt das Projekt *Aladin*[4] [LN05] (ALmost Automatic Data INtegration) an, eine Zusammenarbeit der Humboldt-Universität zu Berlin und des Hasso-Plattner-Institutes für Softwaresystemtechnik in Potsdam. Ziel von Aladin ist es, verschiedene molekularbiologische Datenquellen zu integrieren. Integration bedeutet hierbei dreierlei: der Benutzer soll den gesamten Datenbestand durchsuchen, strukturierte Anfragen stellen und in einer Web-ähnlichen Form durch die Informationen navigieren können.

[1] http://www.rcsb.org/pdb
[2] http://www.cathdb.info
[3] http://scop.mrc-lmb.cam.ac.uk/scop/
[4] http://www.informatik.hu-berlin.de/forschung/gebiete/wbi/research/projects/aladin

Bekannte Ansätze der Datenintegration basieren entweder auf manueller Datenanalyse und -integration durch einen Domänenexperten oder auf automatischer Integration mittels Schema Integration, Schema Mapping und Mediator-Wrapper-Architekturen. Zwar liefert der manuelle, datenzentrierte Ansatz qualitativ gute Ergebnisse; er ist jedoch sehr aufwändig und teuer. Gerade auf Grund der enorm wachsenden Datenmengen sind manuelle Ansätze kaum noch praktikabel [HK04]. Der automatische Ansatz hingegen ist weniger aufwändig bei gleichzeitig geringerer Qualität der Ergebnisse. Er arbeitet schemazentriert, erfordert die Erstellung umfangreicher Schema Mappings und Wrapper und nutzt die Daten selbst nicht [LN05].

Aladin schlägt eine datenzentrierte, fast automatische Integration vor. Gegenüber den oben beschriebenen Verfahren verspricht dies eine hohe Integrationsqualität zu geringen Kosten.

Die genaue Architektur von Aladin kann in [LN05] nachgelesen werden. Unter anderem ist eine Komponente vorgesehen, die Abhängigkeiten zwischen verschiedenen Datenquellen automatisch erkennt. Dies ist von Relevanz, da Biodatenbanken häufig aufeinander verweisen und gerade deshalb ihre Integration interessant ist [HK04].

Diese Abhängigkeiten werden für verschiedene Funktionalitäten von Aladin benötigt. Dazu zählen unter anderem die Entdeckung und Fusion von Duplikaten in den verschiedenen Datenquellen, die Anfragebearbeitung und Suche von Informationen im Gesamtsystem sowie die Visualisierung der Integrationsergebnisse:

- Wenn Datenquellen einander referenzieren, beziehen sie sich meist auf gleiche realweltliche Objekt und enthalten unter Umständen unterschiedliche Informationen darüber. Die Datenbank SCOP etwa enthält Klassifizierungsinformationen über Proteine, die in der PDB beschrieben werden. Diese verschiedenen Informationen über das selbe Objekt zu fusionieren und gesamtheitlich verfügbar zu machen ist eine wichtige Aufgabe von Aladin. Eine Grundvoraussetzung dafür ist es Verweise zwischen den Datenbanken zu kennen.

- Aladin soll Anfragen über Datenbankgrenzen hinweg beantworten können. Für die Komponente, die für die Ausführung der Anfragen verantwortlich sein soll, ist die Kenntnis von Beziehungen zwischen Datenbanken unabdingbar.

- Wie beschrieben, soll der Benutzer auch Web-ähnlich durch die integrierten Daten navigieren können. Interessiert er sich beispielsweise für ein Protein,

das in SCOP kategorisiert ist, soll er per „Link“ auf die ausführliche Beschreibung des Proteins in der PDB gelangen. Dieser Link ist nichts anderes als ein Verweis zwischen Datenquellen.

Die vorliegende Arbeit soll Wege aufzeigen, Beziehungen zwischen Datenquellen automatisiert zu finden. Motiviert ist dies, wie eben dargelegt, durch die Herausforderungen in der Bioinformatik – nichtsdestotrotz kann ein ähnliches Problem auch in anderen Zusammenhängen auftauchen.

1.1 Definitionen

Im Folgenden werden notwendige Begriffe definiert um darauf aufbauend die Aufgabenstellung der vorliegenden Arbeit zu konkretisieren.

Referenzierte und abhängige Attribute Was ist unter den oben erwähnten „Beziehungen“ oder „Verweisen“ zwischen Datenquellen zu verstehen? Im konkreten Fall von molekularbiologischen Datenbanken heißt dies, dass Entitäten in einer Datenquelle auf Entitäten einer anderen Datenquelle verweisen, etwa Krankheiten in einer Datenbank auf beteiligte Proteine in einer anderen Datenbank.

Dies ist technisch meist ähnlich einer Schlüssel-Fremdschlüssel-Beziehung in relationalen Datenbanken umgesetzt. Der eine Entitätstyp besitzt ein Attribut, das die Tupel eindeutig identifiziert. Die Tupel eines anderen Entitätstyps enthalten in einem Attribut einen Verweis auf den gewünschten identifizierenden Wert. Es handelt sich folglich um eine Beziehung zwischen zwei Attributen. Das identifizierende Attribut des ersten Entitätstyps wird als *referenziertes Attribut* bezeichnet, das Attribut des zweiten Entitätstyps als *abhängiges Attribut*. Ein einzelner Wert der Attribute wird entsprechend als *referenzierter Wert* bzw. *abhängiger Wert* bezeichnet.

Wird für zwei Attribute ein solcher Zusammenhang vermutet, welcher erst nachgewiesen werden muss, so werden die beteiligten Attribute in dieser Arbeit als *potenziell abhängiges* und *potenziell referenziertes* Attribut bezeichnet.

Inklusionsabhängigkeiten Eine *Inklusionsabhängigkeit* ist eine konkrete Art von Verweisen zwischen einem abhängigen und einem referenzierten Attribut. Inklusionsabhängigkeiten treten typischerweise innerhalb einer Datenbank in Form von Schlüssel-Fremdschlüssel-Beziehungen auf.