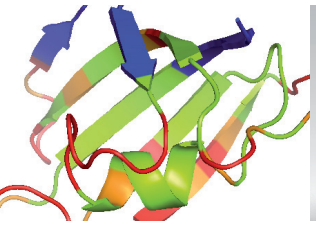
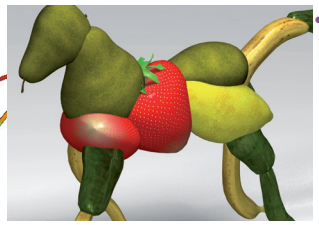
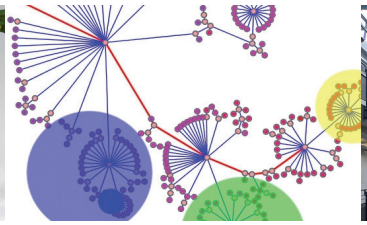
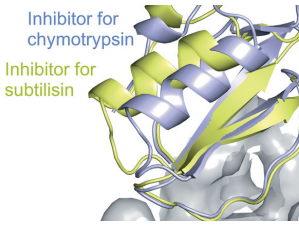




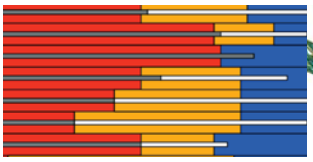
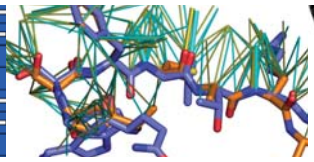


max planck institut
informatik

Bericht 2007/2008

07 08 09 10 11

max planck institut
informatik

Bericht 2007 /2008

Direktorium

Prof. Dr. Thomas Lengauer, Ph.D.
Prof. Dr. Kurt Mehlhorn
Prof. Dr. Hans-Peter Seidel
Prof. Dr. Gerhard Weikum

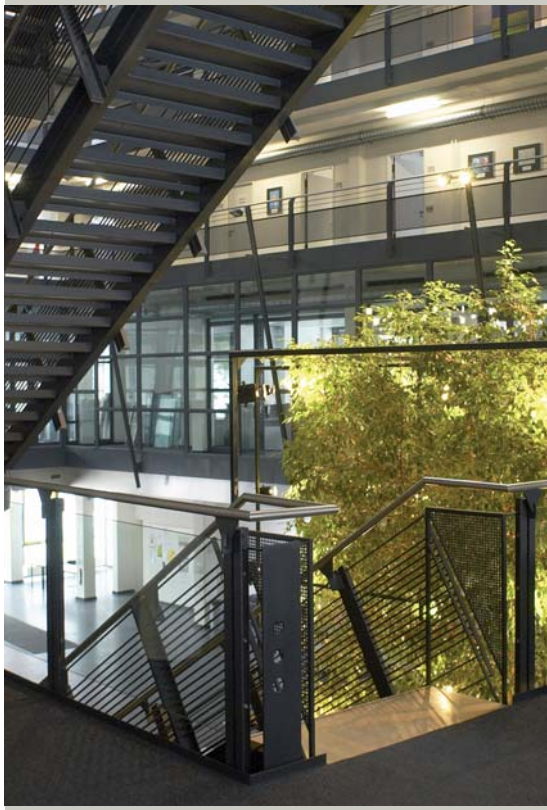
Fachbeirat

Prof. Dr. Pankaj Kumar Agarwal, Department of Computer Science, Duke University, Durham, USA
Prof. Dr. Douglas L. Brutlag, Stanford University, School of Medicine, USA
Prof. Dr. Joseph M. Hellerstein, EECS Computer Science Division, University of California, Berkeley, USA
Prof. Dr. Yannis E. Ioannidis, Department of Informatics & Telecommunications, University of Athens, Griechenland
Prof. Dr. Friedhelm Meyer auf der Heide, Heinz Nixdorf Institut, Universität Paderborn, Deutschland
Prof. Dr. Eugene Myers, Howard Hughes Medical Institute, Ashburn, USA
Prof. Dr. Frank Pfenning, Computer Science Department, Carnegie Mellon University, Pittsburgh, USA
Prof. Dr. Claude Puech, Laboratoire INRIA, Frankreich
Prof. Dr. Éva Tardos, Cornell University, Ithaca, USA
Prof. Dr. Demetri Terzopoulos, Department of Computer Science, University of California, Los Angeles, USA

Kuratorium

Christiane Götz-Sobel, Leiterin der Redaktion Naturwissenschaft und Technik des ZDF, München
Dr. Christoph Hartmann, Minister für Wirtschaft und Wissenschaft, Saarland
Peter Stefan Herbst, Chefredakteur Saarbrücker Zeitung
Prof. Dr. Joachim Hertel, DACOS Software GmbH, Saarbrücken
Prof. Dr. Matthias Jarke, RWTH Aachen
Prof. Dr. Volker Linneweber, Präsident der Universität des Saarlandes
Ministerialdirektor Dr. Wolf-Dieter Lukas, Bundesministerium für Bildung und Forschung
Fritz Raff, Intendant des Saarländischen Rundfunks, Saarbrücken
Dr. Hartmut Raffler, Siemens AG, München
Prof. Dr. Wolffried Stucky, Institut für Angewandte Informatik und Formale Beschreibungsverfahren, Universität Karlsruhe
Dr. Richard Weber, Präsident IHK des Saarlandes, Saarbrücken
Prof. Dr. Margret Wintermantel, Präsidentin der Hochschulrektorenkonferenz

INHALTE



VORWORT

VORWORT

Das Max-Planck-Institut für Informatik legt alle zwei Jahre einen Bericht für die breitere Öffentlichkeit vor. Wir wollen damit allen Wissenschaftsinteressierten Themen, Ziele und Methoden der modernen Informatik und die Arbeiten unseres Instituts vorstellen. Insbesondere hoffen wir, Ihnen, liebe Leser, die Faszination unserer Wissenschaft näher zu bringen.

Das Max-Planck-Institut für Informatik will ein Leuchtturm der Wissenschaft sein. Wir wirken entlang mehrerer Achsen. Erstens durch unsere wissenschaftliche Arbeit, die wir in Publikationen und Büchern aber auch in Form von Software verbreiten. Zweitens durch die Ausbildung von Nachwuchs, insbesondere in der Promotion und danach. Wir produzieren künftige Vordenker und Führungskräfte für Wissenschaft und Wirtschaft. Drittens durch eine Leitrolle im Fach. Wir initiieren und koordinieren große Forschungsprogramme und wir übernehmen Aufgaben in wichtigen Gremien. Viertens als Anziehungspunkt für Talente aus dem In- und Ausland. Von den über 170 Mitarbeiterinnen und Mitarbeitern des Instituts ist etwa die Hälfte aus dem Ausland. Fünftens durch den Transfer unserer Ergebnisse in die Wirtschaft. Dieser Transfer geschieht in Kooperationsprojekten, durch Ausgründungen und durch Personen. Sechstens durch den Aufbau eines Kompetenzzentrums von Weltrang in Kooperation mit den anderen Informatikeinrichtungen am Standort (Fachbereiche Informatik und Computerlinguistik der Universität des Saarlandes, Deutsches Forschungszentrum für künstliche Intelligenz und Max-Planck-Institut für Softwaresysteme). Entlang jeder der Achsen waren wir in den letzten Jahren erfolgreich.

Hervorheben möchten wir den Erfolg des Standorts in der Exzellenzinitiative. Die Informatik des Standorts war in beiden Linien erfolgreich und hat das Cluster Multimodal Computing and Interaction und die Graduiertenschule Informatik eingeworben. Das Institut hat zu beiden Erfolgen einen wesentlichen Beitrag geleistet. Hans-Peter Seidel koordiniert die Arbeit des Exzellenzclusters.

Der Bericht folgt einer einfachen Gliederung. Nach einer Übersicht über Gesamtinstitut, Abteilungen und Forschergruppen stellen wir die Forschungsschwerpunkte des Instituts vor. An diesen Themen werden wir auch in den nächsten beiden Jahren weiter forschen. Im letzten Teil des Berichts finden Sie eine Auswahl von wissenschaftlichen Publikationen für die Jahre 2007 bis 2009 sowie Kennzahlen unseres Instituts für diesen Zeitraum.

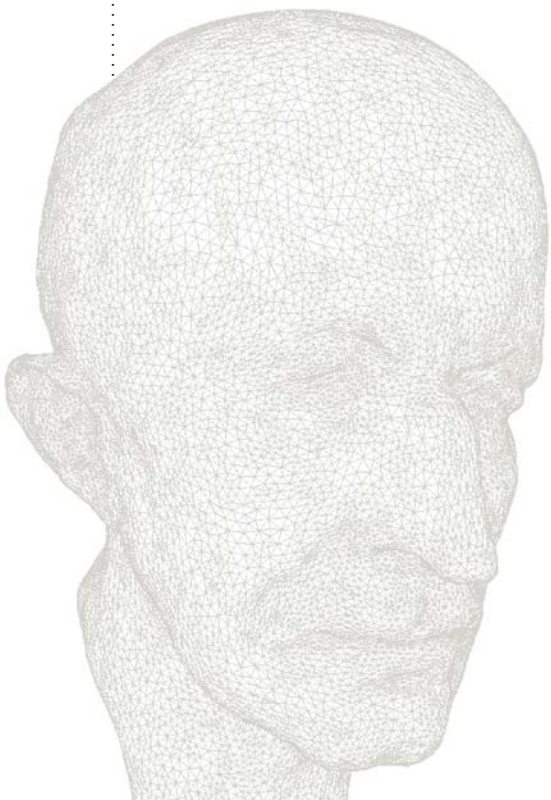
Ich wünsche Ihnen viel Spaß bei der Lektüre dieses Berichts.

Kurt Mehlhorn *Geschäftsführender Direktor*

Das Max-Planck-Institut für Informatik
Ein Überblick

Computersysteme beeinflussen in steigendem Maße unser Leben. Sie bilden die Grundlage nicht nur für praktisch alle geschäftlichen Prozesse, sondern haben schon seit längerem auch in Wissenschaft und Technik und im letzten Jahrzehnt auch in beeindruckendem Maße in unseren Alltag und in unsere Unterhaltung dominant Einzug gehalten. Heute ist die digitale Informationsverarbeitung aus praktisch keinem Bereich des Lebens mehr wegzudenken. Damit ist sie ein gesellschaftlich bestimmender Faktor.

Zusätzlich sind Computer sowie die auf ihnen laufende Software und die aus ihnen gebildeten Netzwerke – allen voran das weltumspannende Internet – wohl die komplexesten Strukturen, die je von Menschenhand geschaffen wurden. In der Tat sind Hardware und in noch weit größerem Maße Software so komplex, dass sie nicht mehr in allen ihren Einzelheiten verstanden werden können. Das macht Computersysteme zu einem sowohl machtvollen als auch mysteriösen Werkzeug. Sowohl das Leistungsvermögen als auch die Geheimnisse von Computersystemen verlangen nach ihrer wissenschaftlichen Erforschung.



ÜBERBLICK

Der wissenschaftliche Umgang mit Computersystemen ist Grundlagenforschung, die jedoch in vielen Fällen in kurzer Zeit zu dramatischen Änderungen des Alltags führt. Gerade die beiden letzten Jahrzehnte machen dies deutlich: World-Wide-Web, Suchmaschinen, Kompressionsverfahren für Video und Musik und sicheres Electronic Banking mittels kryptographischer Methoden sind wenige Jahre nach ihrer Entdeckung in Universitäten und Forschungsinstituten aus unserem Alltag nicht mehr wegzudenken.

Die Max-Planck-Gesellschaft als führende Einrichtung der Grundlagenforschung in Deutschland hat diese Herausforderung angenommen und 1990 das Max-Planck-Institut für Informatik in Saarbrücken gegründet. In 2005 folgte die Gründung des Max-Planck-Instituts für Softwaresysteme mit den Standorten Saarbrücken und Kaiserslautern. In einigen weiteren Instituten gibt es Abteilungen mit starken Informatikkomponenten. Die Bedeutung des Gebiets würde die Gründung weiterer Institute in der Informatik oder in informatiknahen Gebieten rechtfertigen.

Zielsetzung

Das Max-Planck-Institut für Informatik will ein Leuchtturm der Wissenschaft sein. Wir wirken entlang mehrerer Achsen.

Erstens durch unsere wissenschaftliche Arbeit, die wir in Publikationen und Büchern aber auch in Form von Software verbreiten. Zurzeit konzentrieren sich unsere Arbeiten auf Algorithmen für sehr große multi-modale Datenmengen. Multi-modal steht dabei für Text, Sprache, Bilder, Videos, Graphen und hochdimensionale Geometrie.

Zweitens durch die Ausbildung von Nachwuchs, insbesondere in der Promotion und danach. Wir produzieren künftige Vordenker und Führungskräfte für Wissenschaft und Wirtschaft. In unserem Institut arbeiten über 190 Forscher und Forscherinnen, die im Schnitt etwa 3 Jahre bei uns bleiben. Damit stellen wir der Gesellschaft pro Jahr über 60 hervorragend ausgebildete Nachwuchswissenschaftler zur Verfügung.

Drittens durch eine Leitrolle im Fach. Wir initiieren und koordinieren große Forschungsprogramme und wir übernehmen Aufgaben in wichtigen Gremien, etwa dem Wissenschaftsrat. Das Institut hat bei der Einwerbung des Exzellenzclusters Multimodal Computing und Interaction und der Graduiertenschule Informatik durch die Universität des Saarlandes eine wesentliche Rolle gespielt.

Viertens als Anziehungspunkt für Talente aus dem In- und Ausland. Von den Mitarbeiterinnen und Mitarbeitern des Instituts ist etwa die Hälfte aus dem Ausland.

Fünftens durch den Transfer unserer Ergebnisse in die Wirtschaft. Dieser Transfer geschieht in Kooperationsprojekten, durch Ausgründungen und durch Personen. Intel gründete 2009 gemeinsam mit der Universität des Saarlandes Saarbrücken, dem DFKI, und MPI-SWS und MPI-INF das Intel Visual Computing Institute. Intel investiert 12 Millionen US-Dollar in die neue Forschungseinrichtung mit Sitz auf dem Campus der Hochschule. Im Mittelpunkt steht die Entwicklung von zukünftigen Grafik- und Visual Computing-Technologien. Die Investition erfolgt über einen Zeitraum von fünf Jahren und ist bislang Intels umfangreichste Kooperation mit einer Universität in Europa.

Sechstens durch den Aufbau eines Kompetenzzentrums von Weltrang in Kooperation mit den anderen Informatikeinrichtungen am Standort (Fachbereiche Informatik und Computerlinguistik der Universität des Saarlandes, Deutsches Forschungszentrum für künstliche Intelligenz und Max-Planck-Institut für Softwaresysteme).

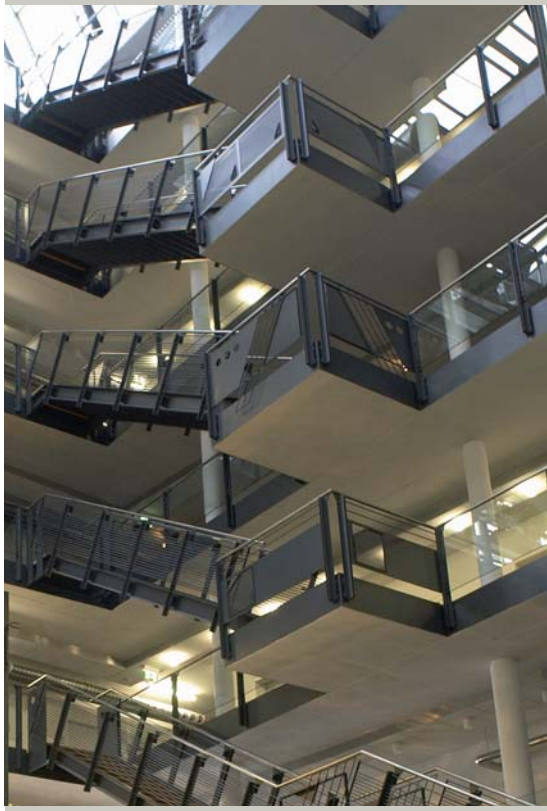
Historie und Institutsstruktur

Das Max-Planck-Institut für Informatik wurde im Jahre 1990 gegründet. Kurt Mehlhorn war der Gründungsdirektor und leitet seitdem am Institut die Abteilung *„Algorithmen und Komplexität“*. Harald Ganzinger war von Anfang an mit dabei und leitete bis zu seinem Tod im Jahr 2004 die Abteilung *„Logik der Programmierung“*, Im Jahre 1999 folgte der Aufbau einer dritten Abteilung *„Computergrafik“* unter der Leitung von Hans-Peter Seidel. Thomas Lengauer kam im Jahre 2001 an das Institut und leitet dort seitdem die Abteilung *„Bioinformatik und Angewandte Algorithmik“*. Seit 2003 leitet Gerhard Weikum am Institut die Abteilung *„Datenbanken und Informationssysteme“*. Die fünfte Abteilung ist in der Besetzung.

Neben den Abteilungen beherbergt das Institut selbständig arbeitende Forschungsgruppen. Derzeit sind die Forschungsgruppen *„Automatisierung der Logik“*, geleitet von Christoph Weidenbach, und *„Informatik für die Genomforschung und Epidemiologie“*, geleitet von Alice McHardy, am Institut tätig.

Forschungsthemen

Der zentrale Forschungsgegenstand des Instituts ist der Algorithmus. Ein Algorithmus ist eine Rechenvorschrift – eine genaue Anweisungsfolge an den Computer, wie er etwas zu berechnen hat. Unsere Arbeitshypothese ist der schnellere Fortschritt in der Informatik durch neue Algorithmen. In den letzten Jahrzehnten stellte die Entwicklung immer schnellerer Rechner einen Meilenstein beim Fortschritt in der Computertechnologie dar. Allerdings wird die dadurch erzielte Beschleunigung der Berechnungen von der Zunahme an Geschwindigkeit, Leistung und Robustheit in den Schatten gestellt, die durch neue Algorithmen erzielt wird. Um ein typisches Beispiel zu nennen: Der Stand der Hardware und Algorithmen im Jahr 1970 ermöglichte die Berechnung einer optimalen Reiseroute für einen Handelsreisenden (ein klassisches Optimierungsproblem und anerkannter Benchmark für die Rechenleistung) durch 120 Städte. Die Erhöhung der Anzahl an Städten von n auf $n+1$ führt zu einem multiplikativen Anstieg der Anzahl an möglichen Routen um einen Faktor n. Legen wir nun also die durch die heutige Technologie höhere Hardware-Geschwindigkeit und die Algorithmen von 1970 zugrunde, so könnten wir lediglich optimale Routen zwischen 135 Städten ermitteln. Es ist der Fortschritt bei den Algorithmen, der es heute ermöglicht, optimale Routen zwischen Tausenden von Städten zu finden. Würden wir uns hier nur auf den Fortschritt bei der Hardware verlassen, wäre eine solche Leistung in Hunderten von Jahren nicht möglich.



Die Aufgabe, die von eigener Hand erstellten Algorithmen und deren Realisierung in Computerprogrammen zu verstehen, ist eine wissenschaftliche und hat zwei wichtige Aspekte. Zum Einen die Frage ob das Programm auch das berechnet was beabsichtigt war und auch nicht „abstürzt“, „einfriert“ oder alle Ressourcen des Computers blockiert. Zum Anderen die Frage, ob das Programm auch „effizient“ ist und der beste mögliche Algorithmus gefunden wurde. In der Abteilung *„Algorithmen und Komplexität“* werden die Ressourcen untersucht, die ein Algorithmus für seine Berechnung braucht. Die wichtigsten Ressourcen sind Rechenzeit *(Wie lange muss ich auf das Berechnungsergebnis warten?)* und Speicherplatz *(Reicht mein Speicher für meine Berechnung?)*. Dabei werden nicht nur neue Algorithmen entwickelt, die den

ÜBERBLICK

Bedarf an Rechenzeit und Speicherplatz minimieren und somit eine direkte hohe praktische Relevanz haben, sondern es werden auch die grundsätzlichen Grenzen dieser Vorgehensweise beleuchtet: Wieviel Rechenzeit/Speicherplatz ist grundsätzlich für eine Berechnung notwendig?

Die Forschungsgruppe *„Automatisierung der Logik"* beschäftigt sich mit Methoden für automatische Beweisverfahren zum Nachweis der korrekten Funktionalität von Computersystemen.

Die Abteilung *„Computergrafik"* widmet sich dem Rechner als Instrument zur Darstellung von Bildern und Filmen. Sie trägt damit der Tatsache Rechnung, dass der Computer zunehmend nicht als Vermittler von Zahlen und Texten sondern vor allem von Bildern und multimedialen Daten in Erscheinung tritt. Auch hier geht es um die Grundfragen: Was ist grundsätzlich machbar? Und: Wieviele Ressourcen werden dafür benötigt? Anstelle des Korrektheitsbegriffs tritt hier der Begriff der naturgetreuen Wiedergabe, ein Konzept, das tiefgehende physikalische Aspekte beinhaltet. Ferner wird der Rechner nicht nur als Bildproduzent eingesetzt, sondern er soll (mithilfe geeigneter Algorithmen) auch Bilder „verstehen", eine Aufgabe, die ebenfalls eine große wissenschaftliche Herausforderung darstellt. Die Abteilung entwickelt auf der anwendungsnahen Seite eine Vielzahl von Verfahren die zur schnellen Erstellung von besseren Bildern und Filmen führen.

Die Abteilung *„Bioinformatik und angewandte Algorithmik"* trägt der Tatsache Rechnung, dass der Computer in den letzten Jahren besonders im Bereich der Lebenswissenschaften eine zentrale Bedeutung erlangt hat, und hier insbesondere bei der Interpretation von biologischen Daten. Der Rechner ist ein wesentliches Instrument der modernen Biologie und Medizin. Das Verständnis biologischer Vorgänge auf molekularer Ebene ist ohne den Rechner nicht möglich, zum einen, weil es in der modernen Biologie immense Datenmengen zu verarbeiten gilt und zum anderen, weil die Komplexität der biochemischen Interaktionen in einem lebenden Organismus das Studium dieser Kreisläufe ohne Zuhilfenahme des Rechners aussichtslos macht. Bioinformatische Methoden sind somit ein Grundbestandteil für die moderne Forschung zur Diagnose und Therapie von Krankheiten.

Die unabhängige Forschungsgruppe für *„Informatik für die Genomforschung und Epidemiologie"* entwickelt unter der Leitung von Dr. Alice McHardy neue Methoden für die Analyse genomischer Sequenzen für Fragestellungen von medizinischer und biotechnologischer Relevanz.

Die Abteilung *„Datenbanken und Informationssysteme"* schließlich widmet sich besonders der Thematik der Verteilung, Organisation und Suche von Daten in großen Computernetzen wie dem Internet. Dabei stehen Aspekte der effektiven Suche nach Information in Netzen (Suchmaschinen wie Google sind entsprechende Instrumente), der Ausfallsicherheit von Methoden im Falle, dass Teile des Netzes nicht zugänglich sind, sowie der effektiven Verteilung von Rechenaufgaben auf im Netz zur Verfügung stehende Rechenleistung (z.B. in Peer-to-Peer-Systemen) im Vordergrund. Der praktische Nutzen dieser Forschung drängt sich auf: Wer hat sich nicht schon einmal gewünscht, mit graphischer statt textueller Information nach Bildern suchen zu können oder selbst bei schwierigen Anfragen von der Suchmaschine auch tatsächlich die relevanten Hits als erste präsentiert zu bekommen?

Exzellenzcluster Multimodal Computing and Interaction

Das Institut spielt eine wichtige Rolle im Exzellenzcluster Multimodal Computing und Interaction. Alle vier Direktoren des Instituts gehören zu den 13 PIs (Principal Investigators) des Clusters und Hans-Peter Seidel koordiniert die Arbeit des Clusters. Wir zitieren die Zusammenfassung des Antrags.

Die letzten drei Jahrzehnte haben dramatische Veränderungen unserer Lebens- und Arbeitsumstände mit sich gebracht, die gemeinhin als Aufbruch in die Informationsgesellschaft beschrieben werden. Technologische Grundlage dieser Umwälzungen sind moderne Computer, die Daten kompakter, billiger und schneller speichern, verarbeiten und übertragen als je zuvor. Diese Kombination gestiegener Leistungsfähigkeit und fallender Preise für Informations- und Kommunikationstechnologie ist beispiellos.

Vor fünf bis zehn Jahren machte Text den Großteil der Inhalte im Web und in anderen Informationssystemen aus. Die starke Verbreitung von Multimedia-Geräten und die verbesserte Medienunterstützung moderner Computer haben diese Beschränkung in den letzten Jahren beseitigt. Digitale Inhalte erscheinen heute in einer Vielzahl von Modalitäten wie Sprache, Bilder, Filme, 3D-Modellen und strukturierten Datensammlungen. Diese Vielfalt wird mit der weiteren Verbreitung drahtloser Kommunikation und moderner Ad-hoc-Sensornetze noch weiter zunehmen. Durch das Internet und das Angebot von digitalen Inhalten über die verschiedensten Kommunikationswege sind Informationen

heute nahezu überall verfügbar. Die Herausforderung besteht nicht länger darin, Informationen zu übermitteln, sondern sie in allen ihren Formen effizient, intelligent und robust zu durchsuchen, zu verstehen und zu organisieren, und Systeme zu schaffen, mit denen das auf natürliche und intuitive Art möglich ist.

Das Exzellenzcluster *Multimodal Computing and Interaction* stellt sich diesen Herausforderungen. Dabei bezieht sich der Begriff *„multimodal"* auf die menschlichen Sinne (besonders Sehen und Hören), die Vielfalt menschlicher Ausdrucksformen und die unterschiedlichen Typen digitaler Daten (wie Text, Sprache, Bilder, Filme, 3D-Modelle). Das Cluster gliedert sich in neun Forschungsbereiche. Vier davon (Text and Speech Processing, Visual Computing, Algorithmic Foundations, Secure Autonomous Networked Systems) sind grundlagenorientiert. Die anderen fünf (Open Science Web, Information Processing in the Life Sciences, Large-Scale Virtual Environments, Synthetic Virtual Characters, Multimodal Dialogue Systems) orientieren sich an den Anwendungen.

Das vorgeschlagene Forschungsprogramm baut auf bestehenden Stärken auf. Die führenden deutschen Fachbereiche für Informatik (UdS-CS) sowie für Computerlinguistik und Phonetik (UdS-CL) der Universität des Saarlandes (UdS) tragen diesen Antrag gemeinsam mit dem Max-Planck-Institut für Informatik (MPI-INF), dem Deutschen Forschungszentrum für Künstliche Intelligenz (DFKI) und dem neu gegründeten Max-Planck-Institut für Softwaresysteme (MPI-SWS). Alle Partner geben ihm höchste Priorität. Die maßgeblich beteiligten Wissenschaftler sind weltweit führend in den zentralen Forschungsbereichen des Clusters. Sowohl die Universität als auch die saarländische Landesregierung haben ihre uneingeschränkte Unterstützung für das Cluster zugesagt.

Diese enge Kooperation und Integration ist ein Garant für die national und international herausragende Stellung des Clusters. Darüberhinaus stellt es eine Zielvorgabe für alle beteiligten Gruppen und Institutionen dar und fokussiert ihre jeweiligen Forschungsprogramme. Obwohl einzelne Gruppen bereits langjährig underfolgreich zusammenarbeiten, ist dies das erste Mal, dass führende Forscher von allen teilnehmenden Institutionen ein gemeinsames Forschungsprogramm vorlegen.

Ein besonderes Ziel des Clusters ist die Qualifikation und Förderung des wissenschaftlichen Nachwuchses. Auf dem Gebiet des Clusters nimmt Saarbrücken hier schon lange eine führende Rolle ein und hat sich über die Jahre den Ruf einer *„Kaderschmiede"* für junge Wissenschaftler erworben. So soll auch jetzt der überwiegende Teil der finanziellen Mittel für die Einrichtung von Nachwuchsgruppen eingesetzt werden. Dabei wird die Hälfte aller Gruppenleiter von der Universität und den teilnehmenden Instituten finanziert. Diese Finanzierung ist dauerhaft über die Laufzeit des Clusters hinaus gesichert. Eine neue Professur für Image, Video and Multimedia Systems mit einer Ausrichtung ähnlich der des Instituts von Prof. Girod in Electrical Engineering an der Stanford University soll anfänglich aus dem Cluster finanziert und anschließend von der Universität dauerhaft übernommen werden. Alle Institute haben offene Stellen in Leitungspositionen, die sie im Forschungsbereich des Clusters besetzen werden.

Die Förderung junger Wissenschaftler zusammen mit der gesicherten lang-

ÜBERBLICK

fristigen Finanzierung der entstehenden Forschungsstrukturen verstärkt die bereits bestehende Exzellenz des Standorts Saarbrücken und etabliert ihn als international führendes Zentrum für Informatik und Sprachtechnologie.

Publikationen und Software

Wir verbreiten unsere wissenschaftlichen Ergebnisse durch Vorträge, Publikationen, in Form von Software und durch Webservices. Unsere Publikationen erscheinen auf den besten Tagungen und in den besten Zeitschriften des Gebiets. Die meisten Publikationen sind auch im Repositorium des Instituts frei zugänglich. Einen Teil unserer Ergebnisse stellen wir auch in Form von Software oder als Webservice zur Verfügung. Bespiele sind CGAL (Computational Geometry Algorithms Library) bzw. der Webservice Geno2pheno zur Beratung bei der HIV Therapie. Veröffentlichungen in Form von Software und Webdiensten machen unsere Ergebnisse direkter und für einen weiteren Kreis von Nutzern zugänglich als klassische Publikationen.

Nachwuchsförderung

Ein weiteres Ziel des Instituts ist die Schaffung eines stimulierenden Klimas für Nachwuchsforscher, damit diese die Möglichkeit haben, ihre eigenen Ideen zu entwickeln und eigene Gruppen aufzubauen. Das Max-Planck-Institut für Informatik betreibt ein aktives Förderprogramm für Doktoranden und Postdoktoranden. Dieses beginnt bis zur Promotion mit dem Doktorandenprogramm der *„International Max Planck Research School for Computer Science“ (IMPRS-CS)* und erlaubt nach der Promotion über internationale Kooperationsabkommen wie dem *„Max Planck Center for Visual Computing“* im Bereich der Computergrafik und die Beteiligung an internationalen Forschungsprojekten den Austausch mit Spitzeninstitutionen in der ganzen Welt. Wir ermutigen damit unsere Nachwuchsforscher, ihre eigenen Forschungsprogramme zu etablieren und zu anderen Einrichtungen zu wechseln. Seit Gründung des Instituts gingen zahlreiche Forscher vom Saarbrücker Max-Planck-Institut für Informatik zu anderen Forschungseinrichtungen und viele von ihnen nahmen eine Professur an.

Gliederung des Berichts

Nach einer Kurzvorstellung der Abteilungen und Forschungsgruppen des Instituts gibt dieser Bericht einen Überblick über die Institutsarbeit, der nach Themenbereichen gegliedert ist. Der Bericht endet mit der Vorstellung der IMPRS-CS, einer Darstellung des Instituts in Zahlen, infrastruktureller Aspekte des Instituts sowie der tabellarischen Auflistung von Kooperationen und Publikationen. Wir wünschen Ihnen viel Freude bei der Lektüre und sind gerne bereit, weiterführende Fragen zu beantworten. Ansprechpartner werden für jedes Thema separat genannt. :::

Algorithmen und Komplexität

PROF. DR. KURT MEHLHORN

ABT. 1

Die Arbeitsgruppe existiert seit Gründung des Instituts und umfasst derzeit etwa 35 Mitarbeiter und Doktoranden. Unsere Ziele sind:

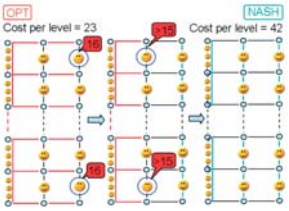
- **herausragende Grundlagenforschung im Bereich Algorithmen,**
- **Umsetzung unserer Grundlagenarbeiten in Demonstratoren und allgemein nützliche Softwarebibliotheken,**
- **Förderung des wissenschaftlichen Nachwuchses in einer stimulierenden Arbeitsgruppe.**

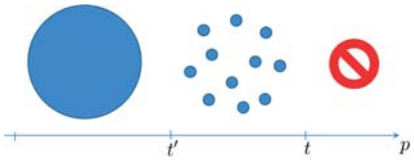
Wir sind in allen drei Aspekten erfolgreich und wirken durch Veröffentlichungen, Software und Personen. Wir publizieren reichlich in den besten Zeitschriften, wir präsentieren unsere Ergebnisse auf den großen internationalen Tagungen des Gebiets, unsere Softwarebibliotheken LEDA und CGAL werden weltweit genutzt, die CompleteSearch Suchmaschine bietet neuartige Möglichkeiten für die effiziente und intelligente Suche in großen Datenmengen. Viele ehemalige Mitglieder der Gruppe haben gehobene Stellen in Forschung und Industrie im In- und Ausland.

KONTAKT

Algorithmen und Komplexität
Sekretariat
Ingrid Finkler-Paul | Christina Fries
Telefon +49 681 9325-100
Email infi@mpi-inf.mpg.de
chfries@mpi-inf.mpg.de

ABT. 1





Algorithmen sind das Herz aller Softwaresysteme. Wir bearbeiten den Entwurf und die Analyse von Algorithmen in vielen Facetten: kombinatorische, geometrische und algebraische Algorithmen, Datenstrukturen und Suchverfahren, verschiedenste Rechnermodelle (sequentiell, parallel, verteilt, flacher Speicher oder Speicherhierarchie), exakte und approximative Lösungen, problem-spezifische Methoden und allgemeine Heuristiken, deterministische und randomisierte Lösungen, obere und untere Schranken, Analyse im schlechtesten Fall und im Mittel. Dabei geht es um die Entwicklung effizienter Algorithmen sowohl für Modellprobleme, d.h., die abstrahierte Version von Anwendungsproblemen, als auch für konkrete Anwendungen, z.B. die intelligente Suche in einer großen Literaturdatenbank. Einen Teil unserer theoretischen Einsichten setzen wir um in Software-Demonstratoren und Softwarebibliotheken.

Herausragende theoretische Ergebnisse der letzten beiden Jahre sind neue Algorithmen zur Lösung verschiedener Probleme auf algebraischen Flächen, zur approximativen Lösung von zweidimensionalen Packungsproblemen, zum Berechnen von gleichmäßig verteilten Punktmengen oder zum Verbreiten von Nachrichten in Netzwerken. Grundlegendes Verständnis für viele algorithmische Fragestellungen liefern unsere jüngsten Analysen von zufälligen planaren Graphen, eine neue Konstruktion von so genannten Epsilonnetzen oder die neuartige Analyse der „least-recently-used“-Heuristic.

Herausragende praktische Ergebnisse der letzten beiden Jahre sind unsere Beiträge zur Softwarebibliothek CGAL, die die Behandlung auch von nicht-linearen Objekten ermöglichen, die CompleteSearch Suchmaschine und ihre Anwendung in einer der wichtigsten Literaturdatenbanken in der Informatik, sowie Algorithmen zur Ansteuerung von OLED-Displays, die die Leuchtdioden einer geringeren Belastung aussetzen und so deren Lebensdauer erhöhen.

Unsere theoretischen und praktischen Arbeiten befruchten sich gegenseitig. Unsere theoretischen Arbeiten sind die Grundlage für die Demonstratoren und Bibliotheken. So beruht die CompleteSearch Engine etwa auf einer neuen Indexstruktur, die mächtiger ist als bekannte Strukturen, aber dennoch nicht mehr Platz benötigt. Die Algorithmen in CGAL nutzen ein tiefes theoretisches Verständnis von algebraischen Kurven und Flächen. Um die Algorithmen zur OLED-Ansteuerung einfach genug für eine Implementierung in einem Chip (wie unlängst geschehen) zu gestalten, war es nötig, auf Standardmethoden wie lineare Programmierung zu verzichten und stattdessen selbstentwickelte kombinatorische Algorithmen zu verwenden.

Die Kombination von theoretischer und experimenteller Forschung in der Algorithmik hat sich inzwischen breiter durchgesetzt. Die DFG unterstützt diese Forschungsrichtung in ihrem Schwerpunktprogramm Algorithm Engineering.

Die Gruppe ist in mehrere internationale Projekte eingebunden: das europäische Projekt ACS (Algorithms for Complex Shapes) und das GIF-Projekt Graphenalgorithmen (mit der Universität Tel Aviv). In Deutschland nehmen wir an dem Schwerpunktprogramm Algorithm Engineering mit drei Teilprojekten teil und sind Teil des Transregio-Sonderforschungsbereiches AVACS (Automatic Verification and Analysis of Complex Systems).

Die Förderung des wissenschaftlichen Nachwuchses ist ein integraler Bestandteil unserer Arbeit. Wir halten Vorlesungen an der Universität des Saarlandes, die sich an Studierende, aber auch unsere Doktoranden richten. Zu unserem Ausbildungskonzept gehört auch, dass wir unsere Doktoranden nach erfolgreicher Promotion erst nach einem mindestens einjährigen Aufenthalt an einer Forschungseinrichtung im Ausland weiterbeschäftigen. Die Gesamtheit dieser Maßnahmen hat dazu geführt, dass die Mitglieder der Arbeitsgruppe nach ihrer Tätigkeit am Max-Planck-Institut für Informatik bestens ausgerüstet auf sehr attraktive Positionen in der Industrie, und das nicht beschränkt auf die Forschung, wechseln oder ihre wissenschaftliche Karriere an führenden Universitäten oder Forschungsinstituten im In- und Ausland fortsetzen. :::

Bioinformatik und Angewandte Algorithmik

PROF. DR. THOMAS LENGAUER, PH.D.

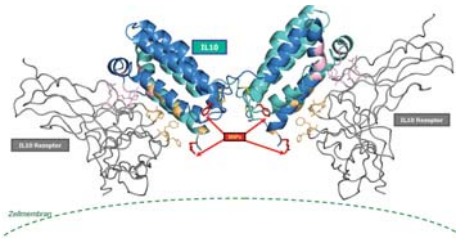
ABT. 3

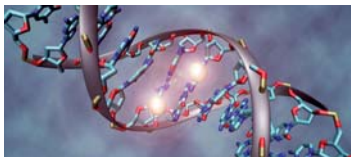
Diese Abteilung existiert seit Oktober 2001 und wird von Prof. Dr. Thomas Lengauer, Ph.D. geleitet. Die Abteilung hat etwa 25 Wissenschaftler. Sie forscht derzeit ausschließlich im Bereich der Bioinformatik.

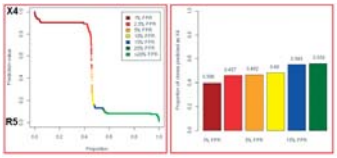
KONTAKT

Bioinformatik und Angewandte Algorithmik
Sekretariat
Ruth Schneppen-Christmann
Telefon +49 681 9325-300
Email ruth@mpi-inf.mpg.de

ABT. 3







Die Abteilung forscht vornehmlich an Themen, die im engeren oder weiteren Sinne für die Diagnose und Therapie von Krankheiten von Belang sind. Auf molekularer Ebene können Krankheitsprozesse auf Anomalien in der „biochemischen Verschaltung“ eines Organismus zurückgeführt werden. Die Bausteine solcher biochemischer Netzwerke sind in der Regel Proteine, die aneinander, an Nukleinsäuren oder an kleine organische Moleküle binden und auf diese Weise chemische Reaktionen katalysieren, das Ablesen von Genen steuern oder Signale innerhalb und zwischen Zellen weiterleiten. Die Aufklärung dieser Funktionsweisen erfordert die Bestimmung der dreidimensionalen Strukturen der beteiligten Proteine, die Analyse von Beziehungen zwischen Proteinstruktur und Proteinfunktion (*„Struktur-Funktionsbeziehungen von Proteinen“*, Seite 28), die Modellierung von Bindungsereignissen zwischen Biomolekülen, sowie die Analyse von komplexen Wechselwirkungs-Netzwerken zwischen Proteinen (*„Analyse menschlicher Proteinnetzwerke“*, Seite 29).

Die Methoden werden darüber hinaus in konkreten Fallbeispielen auf Infektionskrankheiten wie AIDS und auf andere Krankheiten wie Krebs, neurodegenerative und immunologische Krankheiten angewandt. Während bei Krebs die Früherkennung anhand genetischer und so genannter epigenetischer Veränderungen im Vordergrund steht (*„Die genetische Grundlage von Krebsleiden“*, Seite 32, *„Biomedizin 2.0 – Krebsdiagnostik mit dem Computer“*, Seite 33), liegt der Fokus bei anderen Krankheiten auf der Identifizierung und Charakterisierung krankheitsinduzierender Proteine (*„Funktionsanalyse medizinisch relevanter Proteine“*, Seite 30). Bei der Suche nach optimierten Therapien für Infektionskrankheiten spielt AIDS eine herausragende Rolle. Für diese Krankheit gehen wir am Max-Planck-Institut für Informatik sogar noch einen Schritt weiter und analysieren Resistenzen des HI-Virus gegen verabreichte Wirkstofftherapien (*„Analyse von HIV-Resistenzen“*, Seite 31) sowie andere wichtige virale Voraussetzungen für einen auf den Patienten bezogenen effektiven Medikamenteneinsatz (*„Korezeptorvorhersage bei HIV“*, Seite 34).

Ein Großteil der Methodenentwicklung in der Abteilung führt zu Softwaresystemen, die weltweit von zahlreichen akademischen und oft auch industriellen Nutzern angewandt werden. Beispiele hierfür, über die in diesem Band berichtet wird, gibt es im Bereich der Epigenetik (*„Biomedizin 2.0 – Krebsdiagnostik mit dem Computer“*, Seite 33), der Analyse von Proteinfunktion (*„Struktur-Funktionsbeziehungen von Proteinen“*, Seite 28) und Proteinwechselwirkungsnetzen (*„Analyse menschlicher Proteinnetzwerke“*, Seite 29) sowie der Optimierung von AIDS Therapien (*„Analyse von HIV Resistenzen“*, Seite 31, *„Korezeptorvorhersage bei HIV“*, Seite 34).

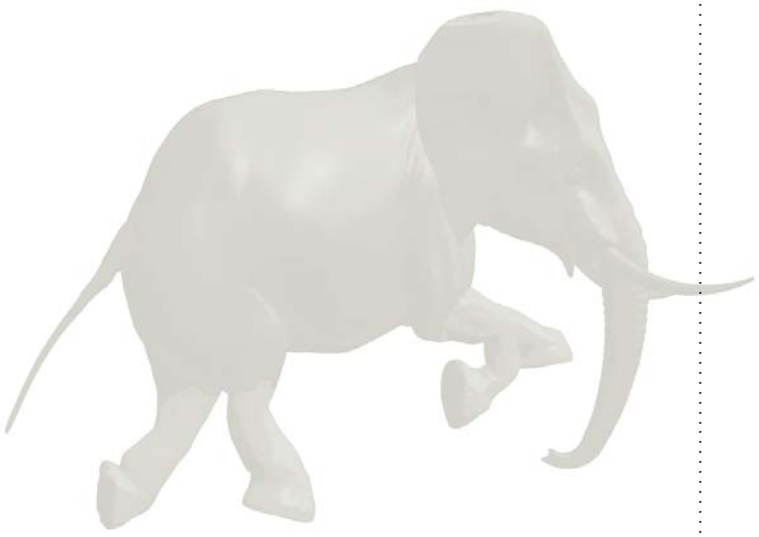
Die Abteilung ist einer der tragenden Säulen des Zentrums für Bioinformatik Saar, einer wissenschaftlichen Einrichtung an der Universität des Saarlandes, die Lehre und Forschung im Bereich der Bioinformatik zum Gegenstand hat. Die Abteilung ist Mitglied des Europäischen Konsortiums „Euresist“, der Klinischen Forschergruppe 129, der Deutschen Forschungsgemeinschaft zur Aufklärung der Funktion des Erregers HCV der Hepatitis C, sowie des vom Bundesforschungsministeriums geförderten Nationalen Genomforschungsnetzes. :::

Computergrafik

PROF. DR. HANS-PETER SEIDEL

ABT. 4

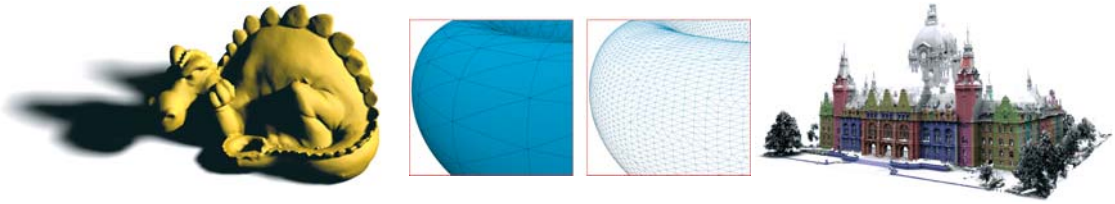
Die Arbeitsgruppe Computergrafik wurde 1999 gegründet und umfasst derzeit knapp 40 Wissenschaftler. Ein wichtiges Charakteristikum der Arbeiten ist die durchgängige Betrachtung der gesamten Verarbeitungskette von der Datenakquisition über die Modellierung bis zur Bildsynthese (3D-Bildanalyse und -synthese). Typisch für das Gebiet ist das Zusammentreffen sehr großer Datensätze mit der Forderung nach schneller, wenn möglich interaktiver, Darstellung.



KONTAKT

Computergrafik
Sekretariat
Sabine Budde
Telefon +49 681 9325-400
Email budde@mpi-inf.mpg.de

ABT. 4



Computer werden heute vielfach dazu benutzt, um Ausschnitte der realen oder einer virtuellen Welt auf dem Rechner nachzubilden, zu simulieren und darzustellen. Aufgrund der Bedeutung visueller Information für den Menschen hat sich die Computergrafik deshalb im vergangenen Jahrzehnt zu einer Schlüsseltechnologie der modernen Informations- und Kommunikationsgesellschaft entwickelt, deren zukünftiges Anwendungspotential durch Schlagworte wie Multimedia, digitales Fernsehen, Telekommunikation, virtuelle Realität oder 3D-Internet lediglich angedeutet ist. Typisch für das Gebiet ist das Zusammentreffen sehr großer Datensätze mit der Forderung nach schneller (wenn möglich interaktiver) visueller Darstellung der Ergebnisse mit hoher Bildqualität. Außerdem soll der Benutzer in die Lage versetzt werden, auf möglichst intuitive Art und Weise mit seiner Umgebung zu interagieren. Hierbei werden verteilte Anwendungen immer wichtiger.

Die genannten Herausforderungen erfordern auch in wissenschaftlicher Hinsicht neue Ansätze. Ein wichtiges Charakteristikum der Arbeitsgruppe ist deshalb die durchgängige Betrachtung der gesamten Verarbeitungskette von der Datenakquisition über die Modellierung (Erzeugung einer geeigneten rechnerinternen Szenebeschreibung) bis zur Bildsynthese (Erzeugung von beliebigen Ansichten). Diese integrierte Sichtweise ist notwendig, um die Leistungsfähigkeit moderner Hardware sowohl bei der Eingabe (bildgebende Verfahren) wie auch bei der Ausgabe (Grafikhardware) adäquat auszunutzen. Inzwischen wurde für diese integrierte Sichtweise der Begriff der 3D-Bildanalyse und -synthese geprägt. Als zentrale wissenschaftliche Herausforderungen ergeben sich hieraus insbesondere die Entwicklung geeigneter Modellierungswerkzeuge zur effizienten Handhabung und Weiterverarbeitung der Datenflut auf der Eingabeseite sowie die Entwicklung neuer Algorithmen zur schnellen und dabei qualitativ hochwertigen Darstellung unter enger Verzahnung mit den Möglichkeiten und Perspektiven moderner Grafikhardware auf der Ausgabeseite.

Die wissenschaftlichen Aktivitäten der Arbeitsgruppe Computergrafik sind in eine Reihe von Projektaktivitäten auf nationaler, europäischer und internationaler Ebene eingebettet.

Von besonderer Bedeutung ist das von der Max-Planck-Gesellschaft und Stanford University mit Unterstützung des BMBF im Oktober 2003 gemeinsam eingerichtete *„Max Planck Center for Visual Computing and Communication“*. Ziel dieses Brückenschlags zwischen den beiden herausragenden Standorten in Deutschland und in den USA ist es, die Forschungsanstrengungen auf diesem Schlüsselgebiet der modernen Informations- und Kommunikationstechnologie zu stärken und zu bündeln, und durch die Etablierung neuer Austauschmechanismen mit attraktiven Rückkehrmöglichkeiten einen wesentlichen Beitrag zur Herausbildung und Rückgewinnung hervorragender Nachwuchswissenschaftler zu leisten. Die gemeinsame Leitung des Zentrums liegt in den Händen von Professor Bernd Girod (Stanford University) und Professor Hans-Peter Seidel (Max-Planck-Institut für Informatik).

Außerdem ist die Gruppe maßgeblich in die Aktivitäten des Exzellenzclusters *„Multimodal Computing and Interaction“* eingebunden. Der Exzellenzcluster wurde im Jahr 2007 im Rahmen der Exzellenzinitiative des Bundes und der Länder neu eingerichtet. Wissenschaftlicher Koordinator des Exzellenzclusters ist Prof. Hans-Peter Seidel.

Eine weitere wichtige Entwicklung ist die Gründung des *„Intel Visual Computing Institute“* im Mai des Jahres 2009. Das neue Forschungsinstitut ist auf dem Campus angesiedelt und wird gemeinsam von Intel, der Universität des Saarlandes, dem DFKI, dem Max-Planck-Institut für Informatik und dem Max-Planck-Institut für Softwaresysteme getragen. Im Governance Board des Instituts ist die Max-Planck-Gesellschaft durch Prof. Seidel vertreten.

Während der vergangenen zehn Jahre haben mehr als 20 ehemalige Nachwuchswissenschaftler der Gruppe Rufe auf Professuren im In- und Ausland erhalten. Die Gruppe hat eine Reihe von Preisen angezogen, darunter neben Nachwuchspreisen für die Wissenschaftler auch den Leibniz-Preis der Deutschen Forschungsgemeinschaft für Professor Seidel. :::

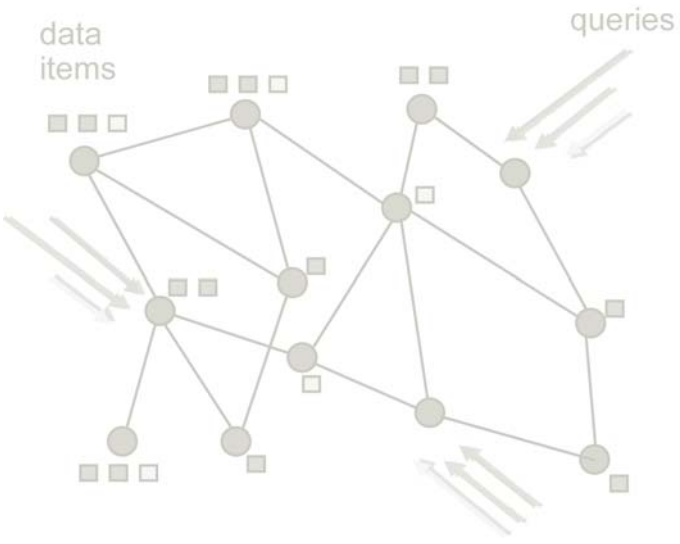
Datenbanken und Informationssysteme

PROF. DR.-ING. GERHARD WEIKUM

ABT. 5

Die von Gerhard Weikum geleitete Abteilung forscht in fünf Themenfeldern:

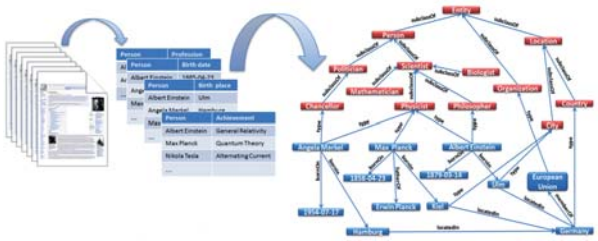
1. **Wissenserschließung im Web mit statistisch und logisch basierten Methoden der automatischen Faktenextraktion aus Internet-Quellen wie Wikipedia**
2. **Text-Mining zur automatischen Klassifikation von Dokumenten und zur Identifikation interessanter Muster in großen Textkorpora, insbesondere in Text- und Web-Archiven über lange Zeitskalen**
3. **Ranking- und Inferenz-Verfahren für Anfragen, bei denen nur die Top-k-Antworten wichtig sind, und für den Umgang mit unsicheren Daten (z.B. für automatisch aus Texten extrahierte Relationen)**
4. **Anfrageverarbeitung und Optimierung von Ausführungsplänen für die effiziente Suche auf strukturierten und semistrukturierten Daten (z.B. im XML- oder RDF-Format)**
5. **Analyse von verteilten Daten, insbesondere in skalierbaren Peer-to-Peer-Systemen, und von Online-Communities, bespielsweise in sozialen Netzen und Web 2.0-Medien.**



KONTAKT

Datenbanken und Informationssysteme
Sekretariat
Petra Schaaf
Telefon +49 681 9325-500
Email schaaf@mpi-inf.mpg.de

ABT. 5



Ein zentrales Leitthema in den wissenschaftlichen Arbeiten der Abteilung ist die automatische Erschließung umfassender Wissensbasen aus Informationsquellen im World Wide Web sowie das Entdecken, Verfolgen und Analysieren von Mustern und individuellen Entitäten (Personen, Organisationen, usw.) und deren Querbeziehungen in dynamischen Web-Quellen. Im YAGO-NAGA-Projekt wurde mittels Faktenextraktion aus Wikipedia und Integration mit der WordNet-Taxonomie eine sehr große Wissenskollektion namens YAGO *(Yet Another Great Ontology)* erstellt. Für die Exploration und intelligente Suche auf YAGO wurde eine neuartige Suchmaschine namens NAGA *(Not Another Google Answer)* entwickelt. Beide beruhen auf dem Semantic-Web-Datenmodell RDF *(Resource Description Framework)*, und die Gruppe hat eine der schnellsten RDF-Suchmaschinen entwickelt, genannt RDF-3X *(RDF Triple Express)*. Eine Vielzahl weiterer Projekte, etwa zur Erschließung multilingualer Information, zur systematischen Sammlung multimodaler Daten (z.B. Photos von Personen), zur Analyse der zeitlichen Veränderungen im Wissen oder zur Deduktion neuer Zusammenhänge auf der Grundlage von wahrscheinlichem, aber nicht hundertprozentig sicherem Wissen, sind um den YAGO-NAGA-Kern gruppiert und miteinander verzahnt.

Die Vision, die diese Arbeiten treibt, ist das langfristig erwartete Zusammenwachsen des *Semantic Web* mit formalen Ontologien und logikorientierter Suche und Inferenz, des *Social Web* (Web 2.0) mit seiner latenten *„Wisdom of the Crowds"*, und des de facto vorherrschenden *Statistical Web*, bei dem die Faktenextraktion aus natürlichsprachigen Texten statistische Lernverfahren benötigt und Suchmaschinen inhärent probabilistisch arbeiten. Das Web könnte damit die Grundlagen einer allumfassenden Wissensbasis werden, die das gesamte Wissen der Menschheit – von vollständigen Enzyklopädien bis zu aktuellen Nachrichten – in formal strukturierter, maschinenlesbarer und damit für Programme und Web-Dienste leicht zu verarbeitender Form enthält. Der Nutzen einer solchen Wissensbasis wäre enorm.

Die Gruppe ist weltweit führend in ihrer Methodik der geschickten Verknüpfung von logikbasierten Algorithmen für strukturierte Datensätze und statistisch basierten Verfahren für unstrukturierte Textdaten. Erstere fallen in das Gebiet der Datenbanksysteme *(DB)* und Datenanalytik, letztere in den Bereich von Information-Retrieval *(IR)* und Suchmaschinen. Historisch sind diese beiden Richtungen getrennt gewesen; ihre Verbindung wird aber bei mehr und mehr gemischten Datenformen immer wichtiger für digitale Bibliotheken, soziale Online-Communities, e-Science-Verbünde und nicht zuletzt in Unternehmen und im Web selbst. In diesem – oft als DB-IR-Integration bezeichneten – aktuellen Forschungsgebiet gehört die Gruppe zu den Trendsettern.

Die Methodik umfasst das gesamte Spektrum von der Theoriebildung bis zum praktischen Einsatz neuer Konzepte in realen Anwendungen und Experimenten. Viele der in der Gruppe entwickelten Prototypsysteme sind als Open-Source-Software öffentlich verfügbar und werden weltweit von anderen Forschungsgruppen genutzt. Dazu gehören insbesondere:

1. das Peer-to-Peer-System *Minerva*, das im EU-Projekt SAPIR zur Suche auf audiovisuellen Inhalten verwendet wird,

2. die XML-Suchmaschine *TopX*, die über mehrere Jahre Spitzenplätze in der INEX-Benchmarking-Reihe erzielt hat und derzeit als Referenzsystem für den INEX-Wettbewerb auf einem semantisch annotierten Wikipedia-Korpus dient,

3. die RDF-Suchmaschine *RDF-3X*, mit der Semantic-Web-Daten und andere graphstrukturierte Daten sehr effizient nach komplexen Mustern durchsucht werden können, und

4. die Softwarewerkzeuge, die zur automatischen Erstellung und Pflege der YAGO-Wissensbasis dienen, sowie die Wissenskollektion YAGO selbst.

Die Abteilung ist an einer Reihe von Drittmittelprojekten beteiligt, insbesondere an den aktuellen EU-Forschungsprojekten *„Living Web Archives"* und *„Living Knowledge"* sowie am DFG-Excellence-Cluster *„Multimodal Computing and Interaction"*. Mehrere Mitarbeiter haben nationale und internationale Dissertationspreise gewonnen. Die wissenschaftlichen Leistungen des Abteilungsleiters, Gerhard Weikum, wurden durch seine Ernennung zum Mitglied der Deutschen Akademie der Technikwissenschaften (acatech) gewürdigt. :::

FG. 1

Automatisierung der Logik

PROF. DR. CHRISTOPH WEIDENBACH

Die unabhängige Forschungsgruppe *Automatisierung der Logik* unter der Leitung von Prof. Dr. Christoph Weidenbach beschäftigt sich mit der kompletten Pipeline von der Erforschung neuer Logiken, bis hin zum automatischen Rechnen in Formeln dieser Logiken.

Damit die Informationstechnologie auch in Zukunft weiter Innovation vorantreiben kann, müssen ihre Produkte (komplexe Systeme, Hardware, Software) noch robuster und qualitativ hochwertiger werden. Sonst sind Anwendungsszenarien, bei denen z.B. Fahrzeuge selbsttätig Bremsmanöver koordinieren, um eine Kollision zu vermeiden, nicht zufriedenstellend realisierbar.

Ein Beitrag zu robusten und qualitativ hochwertigen Systemen (Software und Hardware) der Informationstechnologie ist die Unterstützung ihres Lebenszyklus (Spezifikation, Programmierung, Testen, Wartung) durch die Formalisierung in geeigneten Logiken. Das sind Sprachen mit einer exakten Bedeutung mit deren Hilfe sich die Systeme beschreiben und dann deren Eigenschaften möglichst automatisch nachweisen lassen. Das hört sich utopisch an, ist aber im Kleinen heute in der Software- und Hardwareentwicklung schon lange Realität. So ist es heute Standard, dass Compiler Programme automatisch überprüfen und Programmierfehler wie die Verletzung von Werteschranken in Programmen finden. Dabei ist es egal ob das Programm die Beschreibung eines Hardwarelayouts darstellt oder von dem Prozessor in unserem PC ausgeführt wird.

Abhängig vom Risikopotential des untersuchten Systems (der Software oder Hardware) fordert die Praxis das ganze Spektrum von vollständig mathematisch exakt verifizierten bis hin zu getesteten Systemen. Wir beschäftigen uns mit der mathematisch exakten Analyse von Systemen auf der Basis von Logik. Um hier mit den immer komplexeren Systemen Schritt zu halten, müssen die heute verwendeten formalen Analyseverfahren ihre Produktivität steigern. Das ist das Ziel unserer Forschungsgruppe *„Automatisierung der Logik"*. :::

KONTAKT

Prof. Dr. Christoph Weidenbach
FG. 1 Automatisierung der Logik
Sekretariat Jennifer Müller
Telefon +49 681 9325-900
Email mueller@mpi-inf.mpg.de

UFG. 1

Informatik für die Genomforschung und Epidemiologie

DR. ALICE McHARDY

Seit dem 17. September 2007 arbeitet die unabhängige Forschungsgruppe *Informatik für die Genomforschung und Epidemiologie* unter der Leitung von Dr. Alice McHardy am Max-Planck-Institut für Informatik. In der Gruppe arbeiten zurzeit sieben Wissenschaftler an Verfahren zur Inferenz von Gen-Funktionen, Genotyp-Phänotyp-Relationen und evolutionären Verwandschaften aus Sequenzdaten von viralen, mikrobiellen und eukaryotischen Populationen.

Im Bereich der Metagenomik arbeitet die Gruppe an Verfahren für die Zuordnung genomischer Sequenzfragmente einer mikrobiellen Gemeinschaft zu den sich in der Gemeinschaft befindlichen Organismen. Durch die Analyse mikrobieller Gemeinschaften aus Umgebungen wie dem Magen von Termiten oder dem menschlichen Darm erhofft man neue Einblicke in die Welt der unkultivierbaren Mikroorganismen zu gewinnen. Weiterhin arbeitet die Gruppe an Verfahren zur Funktionszuordnung von Metagenom-Genen mit bisher unbekannter Aufgabe. Hierdurch sollen gezielt Kandidaten für eine detaillierte experimentelle Charakterisierung vorgeschlagen werden können, um so eine Identifizierung von neuen Proteinen mit landwirtschaftlichem, biotechnologischem oder medizinischem Nutzen zu ermöglichen.

Im Bereich der Genomik entwickelt die Gruppe Verfahren zur Inferenz von Genotyp-Phänotyp Beziehungen basierend auf evolutionären Mustern in den Genomsequenzen viraler, mikrobieller und menschlicher Populationen. Diese ermöglichen eine Analyse von genetischen Regionen in Bezug auf deren Relevanz für die Ausprägung eines bestimmten Phänotyps von Interesse. Beispiele hierfür sind die Fähigkeit zur Mensch-zu-Mensch Transmission und die antigenische Evolution von Grippeviren, aber auch genetische Eigenschaften die in Verbindung stehen mit menschlichen komplexen Erkrankungen wie Alzheimer, Morbus Crohn oder Typ 1 Diabetes. :::

KONTAKT

Dr. Alice McHardy
UFG. 1 Informatik für die Genomforschung und Epidemiologie
Sekretariat Ruth Schneppen-Christmann
Telefon +49 681 9325-300
Email ruth@mpi-inf.mpg.de

Max Planck Center for Visual Computing and Communication

PROJEKTLEITUNG: PROF. DR. HANS-PETER SEIDEL

Zur Stärkung des Wissenschafts- und Forschungsstandorts Deutschland wurde im Jahr 2003 das *„Max Planck Center for Visual Computing and Communication"* gegründet. Das Max Planck Center verbindet mit dem Max-Planck-Institut für Informatik in Saarbrücken und der Stanford University zwei weltweit führende Einrichtungen auf dem Gebiet und wurde vom BMBF seit der Einrichtung mit insgesamt 6,9 Millionen Euro gefördert.

Die Forschungsschwerpunkte dieser Kooperation liegen auf der Grundlagenforschung im Bereich des Visual Computing and Communication und umfassen insbesondere die Teilgebiete der Bildaufnahme (Bildakquisition), die Verarbeitung und Analyse von Bilddaten (Bildanalyse), die Erzeugung von Bildern und Bildsequenzen auf der Basis von Aufnahmen oder Simulationen (Bildsynthese), die Visualisierung komplexer Daten sowie den ungestörten und schnellen Austausch von Informationen in komplexen Netzwerken. Dazu bedarf es gleichzeitig der Entwicklung leistungsfähiger Personalcomputer und Betriebssysteme insbesondere als Grafik- und Multimediasysteme.

Stärkung des Wissenschafts- und Forschungsstandorts Deutschland im Bereich der Informatik

Ein wesentliches Ziel des Programms ist die Herausbildung und Förderung des wissenschaftlichen Nachwuchses, indem es besonders qualifizierten jungen Informatikern einen Weg zu frührer wissenschaftlicher Selbständigkeit bei gleichzeitiger enger Einbettung in ein international kompetitives stimulierendes wissenschaftliches Umfeld eröffnet. Hierbei wird besonders herausragenden jungen Postdoktoranden die Möglichkeit gegeben, unter der Betreuung je eines Mentors aus Deutschland und den USA eigenständig mit einer kleinen Arbeitsgruppe bis zu fünf Jahre zu forschen. Nach einem zweijährigen Aufenthalt in Stanford, wo sie den Status eines „Visiting Assistant Professors" innehaben *(Phase I)*, kehren die Wissenschaftler zurück nach Deutschland und setzen Ihre Arbeit als Nachwuchsgruppenleiter am Max-Planck-Institut für Informatik fort *(Phase II)*. Die zweite Phase des Programms steht grundsätzlich auch herausragenden rückkehrwilligen Postdoktoranden aus anderen Ländern offen.

Aktueller Stand

Seit nunmehr sechs Jahren setzt dieses Modell dem oft beobachteten „Brain Drain" in die USA attraktive Perspektiven in Deutschland entgegen und liefert so einen Beitrag zur Herausbildung und Sicherung hochqualifizierten wissenschaftlichen Nachwuchses und damit zur nachhaltigen Stärkung der Innovations- und Wettbewerbsfähigkeit des Standortes. In dieser Zeit hat sich das Max-Planck-Center in der internationalen Fachwelt den Ruf einer echten Talentschmiede erarbeitet. Seit Einrichtung des Programms im Jahr 2003 haben bisher insgesamt 12 Nachwuchswissenschaftler das Programm vollständig durchlaufen. Hiervon wurden inzwischen 11 Nachwuchswissenschaftler auf Professuren berufen, 10 davon in Deutschland, 7 davon auf W3-Professuren.

Darüberhinaus wurden die durch das Programm geförderten Nachwuchswissenschaftler mit einer Reihe von renommierten Nachwuchspreisen im Fachgebiet ausgezeichnet (Eurographics Young Researcher Awards 2005 bzw. 2009, Olympus-Preis 2007, Heinz-Billig-Preis 2007, Otto-Hahn-Medaille der Max-Planck-Gesellschaft 2007, SaarLB Forschungspreis 2008) und erhielten bei führenden internationalen Fachtagungen für ihre dort eingereichten und vorgestellten Arbeiten mehrfach Best-Paper-Awards. Insgesamt entstanden mehr als 450 wissenschaftliche Publikationen.

Der Erfolg des Programms zeigt, dass es in Deutschland durchaus möglich ist, im weltweiten Kampf um die besten Köpfe erfolgreich zu bestehen. Kernelemente dieses erfolgreichen Programms sind dessen internationale Ausrichtung, das flexible und hochdynamische Forschungsprogramm, dessen Ausrichtung die Bewerber selbst wesentlich mitgestalten, die frühe wissenschaftliche Selbständigkeit der Nachwuchswissenschaftler bei gleichzeitiger enger Einbettung in ein international kompetitives, stimulierendes wissenschaftliches Umfeld sowie die attraktiven Rückkehrperspektiven.
So könnten diese strukturellen Elemente des Programms möglicherweise auch für andere Fachdisziplinen Modellcharakter haben. :::

KONTAKT

Prof. Dr. Hans-Peter Seidel
Sekretariat Sabine Budde
Telefon +49 681 9325-400
Email budde@mpi-inf.mpg.de

Forschungsschwerpunkte

BIOINFORMATIK

Die Bioinformatik ist eine Schlüsseldisziplin für den schnelleren Erkenntnisfortschritt in den Biowissenschaften, wie Biotechnologie, Pharmazie und Medizin. Die Bioinformatik vertieft und beschleunigt mit Hilfe des Computers die Planung von höchst komplexen biologischen Experimenten und die Interpretation der in sehr großen Mengen anfallenden Daten.

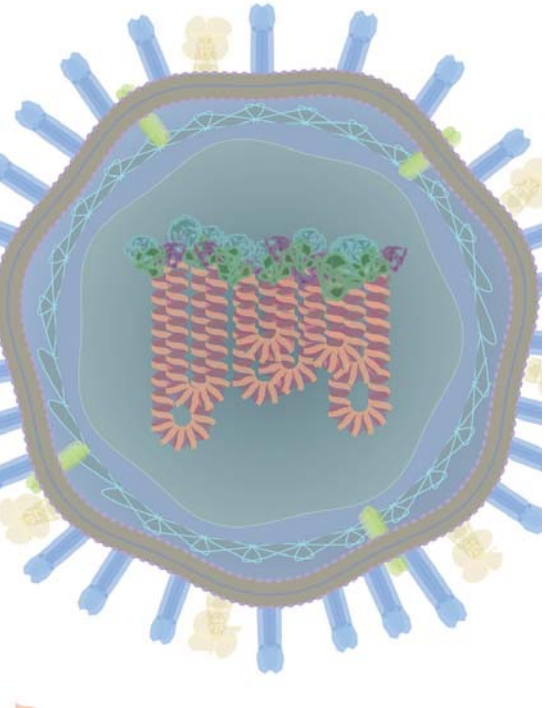
Seit etwa 15 Jahren trägt die Bioinformatik wesentlich zum Erkenntnisgewinn in den Biowissenschaften bei. Sie ist Teil einer Revolution der Biologie. Sie unterstützt Forscher bei der Planung von Experimenten, sie sammelt Daten, die aus allen Bereichen des Organismus stammen und wertet diese Daten mit computergestützten Methoden aus. Mit ihrer Hilfe dringen Wissenschaftler bis zu den molekularen Abläufen in der Körperzelle vor, der Grundeinheit von lebenden Organismen – in ein komplexes Materie, Energie und Information verarbeitendes System, in dem molekulare Prozesse auf vielen verschiedenen Ebenen zusammenwirken. Das Genom speichert den Bauplan der Zelle und den Ablaufplan ihrer Stoffwechselprozesse. Um diese Zellprozesse zu unterhalten, müssen immer wieder Teile des Genoms „abgelesen" werden, so etwa die Gene. Sie enthalten die Baupläne von Proteinen, der zellulären Molekular-Maschinen. Das Ablesen der Gene wiederum wird durch komplexe molekulare Netzwerke gesteuert. Für die Synthese von Proteinen und auch für ihren Abbau gibt es spezielle molekulare Komplexe, die selbst wieder detaillierter molekularer Steuerung unterliegen. Die Zelle wandelt Energie um, sie kommuniziert mit Zellen in ihrer Umgebung, sie nimmt unterschiedliche Strukturen an und bewegt sich. Sie reagiert auf Änderungen in ihrer Umgebung, zum Beispiel auf Veränderungen des Lichts, der Temperatur und des pH-Werts, und sie wehrt Eindringlinge ab. Fehlsteuerungen dieser Prozesse sind die molekulare Grundlage für Krankheiten. Therapien zielen darauf ab, ein verträgliches molekulares Gleichgewicht wieder herzustellen.

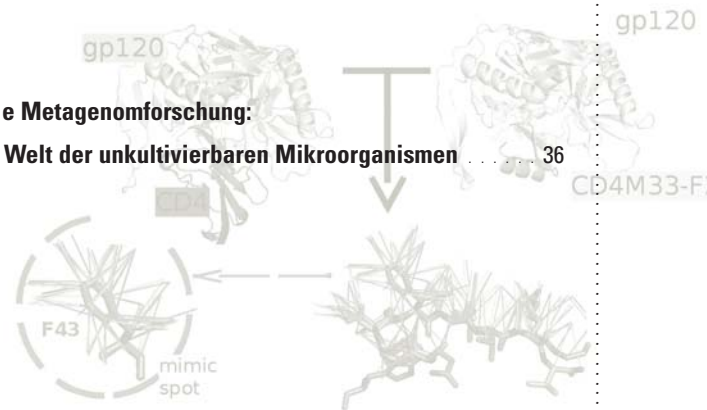
Seit gut zehn Jahren wird die klassische biologische Forschung, die bis dahin zumeist auf sehr eng eingegrenzte Teilsysteme der Zelle konzentriert war, durch Hochdurchsatz-Experimente ergänzt. Diese erfassen zellweit Daten, etwa durch eine umfassende Analyse des Genoms oder durch Messen von Häufigkeiten aller abgelesenen Gene (Transkriptom). Erfasst werden ferner die Varianten der von der Zelle verwandten Proteine (Proteom) und deren Wechselwirkungen (Interaktom). Aus diesen Daten kohärente Einsichten über die Biologie der Zelle, die Grundlagen von Krankheiten sowie Ansätze für Therapien abzuleiten, ist eine hoch komplexe informationstechnische Aufgabe. Dieser stellt sich die Bioinformatik. Am Max-Planck-Institut für Informatik wird in vielen der hier angesprochenen Bereiche geforscht. So erarbeiten die Experten am Institut beispielsweise neue Wege zur Berechnung von auf den einzelnen Patienten abgestimmten, optimalen Wirkstoffkombinationen – beispielsweise zur Behandlung von AIDS (siehe *„Analyse von HIV-Resistenzen"*, Seite 31).

Damit hat die Bioinformatik den hybriden Charakter einer Grundlagenwissenschaft, die frühzeitig klare Anwendungsperspektiven definiert. Diese einzigartige Eigenschaft wird durch eine beträchtliche Zahl von Ausgründungen aus bioinformatischen Forschungsgruppen unterstrichen. So hat beispielsweise Professor Lengauer mit seinen Mitarbeitern die Firma BioSolveIT GmbH gegründet, die Software für den Entwurf von Medikamenten entwickelt und vertreibt. Zu den Nutzern dieser Software gehören weltweit über hundert Pharmafirmen.

Das von der DFG geförderte Zentrum Bioinformatik Saar, dessen Vorsitzender Professor Lengauer derzeit ist, wurde unter den fünf Zentren in Deutschland bei der letzten Bewertung (2007) als führend in der Forschung eingestuft. :::

BEITRÄGE





Struktur-Funktionsbeziehungen von Proteinen

Hintergrund: Proteine, Bausteine des Lebens

Proteine verrichten eine Vielzahl von Aufgaben in Zellen. Die meisten zellulären Prozesse beruhen auf dem Zusammenspiel mehrerer Proteine. Für viele der etwa 25.000 Protein kodierenden Gene im Menschen ist jedoch noch nichts über die Funktion der entsprechenden Proteine bekannt.Ein Protein besteht aus einer langen Kette von Aminosäuren die sich in der Zelle zu einer charakteristischen dreidimensionalen Struktur faltet. Diese 3D-Struktur des Proteins hat direkte Auswirkungen auf seine Funktion. So bestimmt beispielsweise die Gestalt bestimmter Bereiche an der Proteinoberfläche, ob und wie gut das Protein an bestimmte andere Proteine mit komplementärer Oberfläche binden kann. Neben der Interaktion mit anderen Proteinen spielt die Struktur auch für das Zusammenspiel mit kleinen Molekülen wie etwa Wirkstoffen in Medikamenten oder mit DNA eine Rolle. Die molekularen Details haben folglich einen entscheidenden Einfluss auf die Fragen: Wer bindet an wen? Welche Proteinregionen binden aneinander? Wie stark ist eine Bindung? Was passiert mit dem gebundenen Partner: wird er transportiert, gespeichert, chemisch modifiziert? Was ist die Rolle dieser einzelnen Funktion im größeren Kontext zellulärer Prozesse?

Methoden und Verfahren

Struktur-Funktionsbeziehungen von Proteinen sind ein Forschungsschwerpunkt der Abteilung 3 des Max-Planck-Instituts für Informatik. Exemplarisch sollen hier zwei Verfahren vorgestellt werden:
(1) Vorhersage von molekularer Funktion bei gegebener Proteinstruktur.
(2) Vergleich von Bindestellen und Interaktionsmustern in Kontaktstellen zwischen zwei Proteinen, um wiederkehrende Muster und Gesetzmäßigkeiten aufzudecken.

Funktionsvorhersage: GOdot

Mit dem GOdot System kann die molekulare Funktion eines Proteins mit bekannter Struktur vorhergesagt werden. Das geschieht dadurch, dass das neue Protein mit einer Menge von Referenz-Proteinen strukturell verglichen und so in eine lokale Region innerhalb eines mathematischen Ähnlichkeitsraumes eingebettet wird. In diesem Ähnlichkeitsraum wird die lokale Konservierung von molekularen Funktionen analysiert und dadurch lässt sich abschätzen, welche Funktionen, die in der betrachteten Region des Ähnlichkeitsraumes vorherrschen, auf das Zielprotein übertragen werden können. Das GOdot System ist als Webserver verfügbar. Dieser bietet zusätzlich Visualisierungsmöglichkeiten. In obengenannten Ähnlichkeitsraum wird die lokale Umgebung eines Zielproteins mittels Multidimensional Scaling in eine Ebene projiziert und graphisch dargestellt; jedes Protein wird als ein Punkt dargestellt, wobei ähnliche Proteine entsprechend nahe beieinander liegen. Die Proteine lassen sich dann entsprechend ihrer Funktionen und Familienzugehörigkeiten farblich markieren [Abbildung 1].

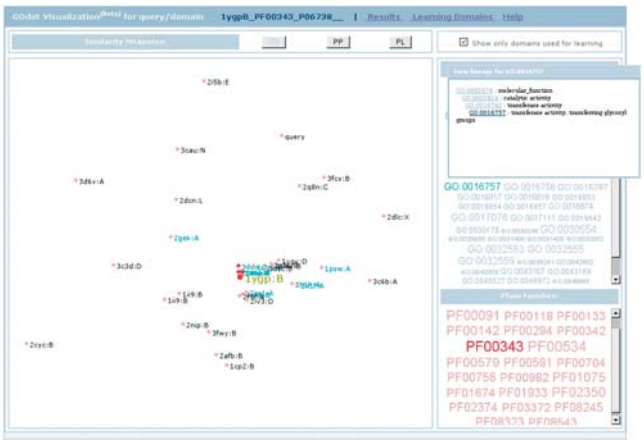


Abbildung 1: Funktionsvorhersage: GOdot

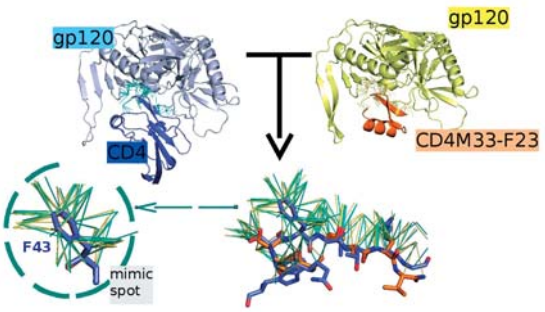


Abbildung 2

Charakterisierung von Protein-Protein Interaktionen: Galinter

Aufgaben wie die Vorhersage von Interaktionspartnern zu Proteinen oder der Entwurf von Protein-Interaktionen oder von kleinen Molekülen, die auf dem Wege einer medikamentösen Therapie Interaktionen blocken sollen, erfordern ein tiefes Verständnis der Funktionsweise von Protein Bindestellen. Hierzu haben wir Galinter, ein neues Programm zum strukturellen Vergleich von Protein-Protein Interaktionen entwickelt. Mit Galinter lassen sich Protein-Protein Interaktionen vergleichen und visualisieren. Damit lässt sich der Bindungsmodus der Interaktionspartner analysieren.

Wir haben mit dieser Methode zum Beispiel den Bindungsmodus des gp120 Hüllproteins von HIV mit dem entworfenen Protein CD4M33-F23 rechnerisch untersucht. Damit HIV Wirtszellen infizieren kann, muss das gp120 Protein von HIV an CD4 Rezeptoren auf der Oberfläche menschlicher Wirtszellen binden. Das CD4M33-F23 Protein wurde als Interaktions-Antagonist entwickelt, indem die CD4 Bindungsstelle auf ein Skorpion Toxin Protein übertragen wurde. Unsere Analyse mit Galinter ergibt, dass der Bindungsmodus von gp120 mit dem entworfenen CD4M33-F23 Protein dem Bindungsmodus von gp120 mit dem CD4 Protein entspricht [Abbildung 2]. :::

KONTAKT

Francisco Silva Domingues
ABT. 3 Bioinformatik und Angewandte Algorithmik
Telefon +49 681 9325-304
Email doming@mpi-inf.mpg.de

Ingolf Sommer
ABT. 3 Bioinformatik und Angewandte Algorithmik
Telefon +49 681 9325-306
Email sommer@mpi-inf.mpg.de
Internet http://godot.mpi-inf.mpg.de

Analyse menschlicher Proteinnetzwerke

Proteininteraktionen in Zellen

Proteine sind an nahezu allen Lebensvorgängen in Zellen beteiligt, beispielsweise am Transport von Stoffen, der Übermittlung von Signalen oder der Beschleunigung biochemischer Prozesse. Dabei interagieren Proteine häufig miteinander, um ihre biologischen Aufgaben zu erfüllen. Um die entsprechenden Prozesse in den Zellen zu verstehen, ist es daher notwendig, das komplexe Zusammenspiel der Proteine zu kennen.

Störungen von Interaktionen zwischen Proteinen stehen darüber hinaus im Verdacht, eine entscheidende Rolle bei der Entstehung von Erkrankungen wie Krebs oder Darmentzündungen zu spielen (siehe auch *„Funktionsanalyse medizinisch relevanter Proteine“,* Seite 30). Durch ein besseres Verständnis der Wechselwirkungen von Proteinen könnten somit auch neue Behandlungsmöglichkeiten eröffnet werden.

Verfügbarkeit von Proteininteraktionsdaten

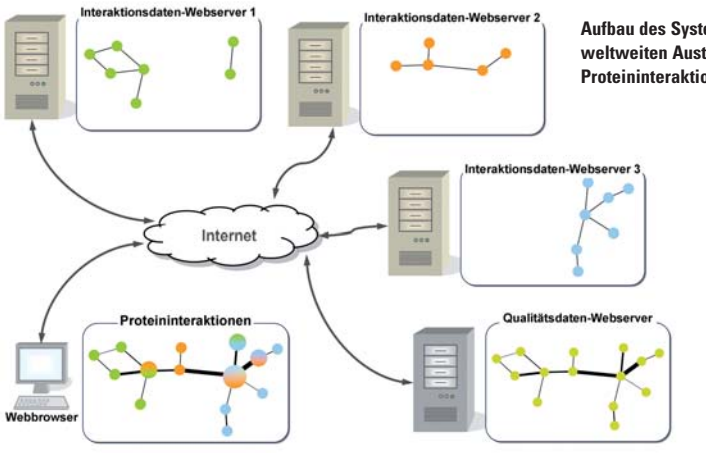
Um das volle Potenzial der weltweit verteilten Informationen zu Proteininteraktionen ausschöpfen zu können, müssen diese Daten für Forscher leicht und jederzeit aktuell verfügbar sein. Gerade in den letzten Jahren hat sich die Forschung auf diesem Gebiet rasant weiter entwickelt, so dass immer mehr Interaktionsdaten durch Forschergruppen auf der ganzen Welt generiert werden. Diese Flut an verfügbaren Informationen ist für die Wissenschaft zwar eine große Chance, für den einzelnen Forscher gestaltet es sich jedoch zunehmend schwierig, den Überblick über die bereits vorhandenen Daten zu wahren. Um alle bekannten Interaktionspartner eines bestimmten Proteins zu finden, muss ein Forscher beispielsweise derzeit bis zu hundert verschiedene Proteininteraktions-Datenbanken im Internet besuchen und ihre Inhalte zusammenführen. Da außerdem bisher erst ein geringer Prozentsatz aller im Menschen stattfindenden Wechselwirkungen zwischen Proteinen bekannt ist, wird sich der damit verbundene Zeitaufwand in naher Zukunft weiter vergrößern.

Neues System für den Datenaustausch

In Zusammenarbeit mit Kollegen in England am Wellcome Trust Sanger Institute und dem European Bioinformatics Institute haben wir daher ein neuartiges System entwickelt, um den Zugang zu Proteininteraktionsdaten stark zu vereinfachen. In der Vergangenheit vorgestellte Ansätze hatten das Ziel, die global verstreuten Informationen durch zentrale Datenspeicherung besser verfügbar zu machen. Die Zusammenführung der Daten verursacht jedoch einen großen zeitlichen Pflegeaufwand, da die zentral gespeicherten Informationen stets auf dem aktuellen Stand zu halten sind. Daher haben wir als Grundlage für unser System eine verteilte Architektur gewählt [Abbildung]. Statt all die Informationen zu Proteinwechselwirkungen an einem zentralen Ort zu sammeln, bleiben sie bei unserem System dort, wo sie herstammen. Das hat insbesondere den Vorteil, dass die angezeigten Interaktionsdaten stets aktuell sind, da die verteilten Datenquellen erst bei Bedarf über das Internet abgefragt werden.

Qualitative Bewertung der Daten

Auch das Problem der Bewertung der unterschiedlichen Qualität von Interaktionsdaten lässt sich mit dem unserem System lösen. Jede bisher angewandte Messmethode zur Bestimmung von Proteininteraktionen hat bestimmte Schwächen. So werden in Experimenten Interaktionen zwischen Proteinen manchmal fälschlicherweise nachgewiesen oder bleiben unentdeckt, obwohl sie in der Zelle vorkommen. Unser System ermöglicht es nun, die Zuverlässigkeit von einzelnen gemessenen Interaktionen zu bewerten. Da es viele verschiedene Bewertungskriterien für die Qualität von Interaktionsdaten gibt, kann keine Institution alle anbieten. Durch eine dezentrale Verteilung der Qualitätsdaten kann sich jede Forschungsgruppe auf einzelne Aspekte der Qualitätsbeurteilung fokussieren, zum Beispiel auf die Bewertung der funktionellen Ähnlichkeit der Interaktionspartner. Die einzelnen Qualitätswerte werden dann von unserem System über das Internet automatisch zusammengetragen und dem Benutzer auf einer Webseite präsentiert. Die freie Verfügbarkeit und die einfache Erweiterbarkeit unseres Systems erlaubt es Forschergruppen, ihre Ergebnisse anderen Wissenschaftern ohne Aufwand weltweit zur Verfügung zu stellen. :::



Aufbau des Systems zum weltweiten Austausch von Proteininteraktionsdaten

KONTAKT

Mario Albrecht
ABT. 3 Bioinformatik und Angewandte Algorithmik
Telefon +49 681 9325-327
Email mario.albrecht@mpi-inf.mpg.de

Hagen Blankenburg
ABT. 3 Bioinformatik und Angewandte Algorithmik
Telefon +49 681 9325-328
Email hagen.blankenburg@mpi-inf.mpg.de
Internet http://medbioinf.mpi-inf.mpg.de

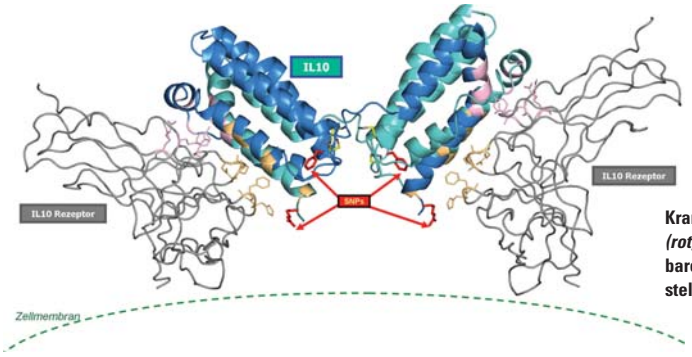
Funktionsanalyse medizinisch relevanter Proteine

Genetische Variation der DNA

Verschiedene Menschen unterscheiden sich durch individuelle genetische Veränderungen voneinander. Am häufigsten treten sogenannte *Single Nucleotide Polymorphism* (SNP) auf. Dabei handelt es sich um Punktmutationen der DNA, also Änderungen von Buchstaben im genetischen Text, die mit größerer Wahrscheinlichkeit als andere Sequenzvariationen auftreten. Populationsstudien zeigen, dass bestimmte SNPs gehäuft in einzelnen Bevölkerungsgruppen auftreten und für die Anfälligkeit gegenüber Krankheiten und die Schwere ihres Verlaufs maßgeblich sein können. Außerdem können SNPs die Verträglichkeit und Wirkung von Medikamenten beeinflussen. Die medizinische Forschung ist daher besonders an den molekularen Veränderungen interessiert, die durch SNPs verursacht werden.

Proteinstruktur und -funktion

Die DNA-Sequenz eines Gens gibt vor, in welcher Reihenfolge Aminosäuren zu einem Protein zusammengesetzt werden. Somit stellt sie den Bauplan eines Proteins dar. Proteine haben eine charakteristische dreidimensionale Struktur, die durch atomare Wechselwirkungen innerhalb der Proteinstruktur erzeugt wird. Die meisten molekularen Prozesse laufen aufgrund von Wechselwirkungen zwischen Proteinen ab (siehe auch „*Analyse menschlicher Proteinnetzwerke*“, Seite 29). Änderungen der Proteinstruktur können diese Interaktion stören. Derartige Modifikationen können zum Beispiel durch Punktmutationen wie SNPs verursacht werden. So kann eine Punktmutation zum Austausch einer Aminosäure im Protein führen und bedeutende Veränderungen der intra- und intermolekularen Wechselwirkungen des Proteins zur Folge haben. Die hierdurch beeinträchtigte Funktion des Proteins und des involvierten biologischen Prozesses kann sogar Krankheiten verursachen. Bekannte Beispiele sind die Sichelzellanämie und die Mukoviszidose.



Krankheitsassoziierte SNPs *(rot)* befinden sich in unmittelbarer Nähe der Rezeptorbindestelle für das Protein IL10.

SNPs bei Darmentzündungen

Klinische Assoziationsstudien, wie sie unsere medizinische Kooperationspartner an der Universitätsklinik Kiel durchführen, durchsuchen die DNA chronisch erkrankter Personen systematisch nach genetischen Gemeinsamkeiten. Durch den Vergleich von Patienten mit einer gesunden Vergleichsgruppe werden krankheitsassoziierte SNPs in der DNA des menschlichen Genoms entdeckt. Bioinformatiker analysieren dann diese Daten mit computergestützten Methoden, um die genaue Funktion relevanter humaner Proteine und ihrer SNPs bei der Entstehung und dem Verlauf von Krankheiten besser zu verstehen. Colitis ulcerosa ist eine chronisch-entzündliche Darmerkrankung mit familiärer Häufung, deren molekulare Ursachen noch weitgehend unbekannt sind. Es wird vermutet, dass die Barrierefunktion des Immunsystems im Darm gestört ist und die Darmwand überempfindlich auf die dort natürlich vorkommenden Bakterien reagiert. Ein Zusammenhang zwischen dieser Krankheit und Variationen im Gen *IL10* konnte kürzlich in einer Assoziationsstudie festgestellt werden. Das Protein IL10 ist ein Botenstoff im Immunsystem, der spezifisch an einen bestimmten Zellrezeptor bindet [Abbildung] und dadurch wichtige Signalwege zur Hemmung von Entzündungen ermöglicht.

Computeranalyse des Proteins IL10

Die möglichen Auswirkungen zweier SNPs auf die Struktur und Funktion des Proteins IL10 wurden anhand bioinformatischer Verfahren näher untersucht. Dabei wurde festgestellt, dass sich beide Punktmutationen in unmittelbarer Nähe der Bindestellen des Zellrezeptors befinden [Abbildung]. Es ist naheliegend, dass diese Sequenzvariationen die Interaktion zwischen IL10 und seinem Rezeptor stören können. Hierdurch werden entzündungshemmende Signalwege unterbunden, was der Entstehung chronischer Erkrankungen Vorschub leistet.
Deswegen wird nun die therapeutische Verabreichung von IL10 an Patienten zur Behandlung von Colitis ulcerosa in Betracht gezogen.

Medizinische Bioinformatik

Auf diese Weise unterstützt die medizinische Bioinformatik klinische Forschungsarbeiten und hilft mit, genetische Befunde, wie hier von Patienten mit Colitis ulcerosa, auf molekularer Ebene zu interpretieren und besser zu verstehen. Dadurch beschleunigen ihre rechnergestützten Methoden die Aufklärung der molekularen Ursache von Erkrankungen und ermöglichen eine schnellere Medikamentenentwicklung. :::

KONTAKT

Mario Albrecht
ABT. 3 Bioinformatik und Angewandte Algorithmik
Telefon +49 681 9325-327
Email mario.albrecht@mpi-inf.mpg.de

Gabriele Mayr
ABT. 3 Bioinformatik und Angewandte Algorithmik
Telefon +49 681 9325-304
Email gabriele.mayr@mpi-inf.mpg.de
Internet http://medbioinf.mpi-inf.mpg.de

Analyse von HIV-Resistenzen

HIV Therapie

Seit der Entdeckung des Humanen Immundefizienz-Virus *(HIV)* vor etwa 25 Jahren, dem Auslöser von AIDS, sind Medikamente entwickelt worden, die in vier verschiedenen Phasen des viralen Lebenszyklus eingreifen:

1 Eintritt in die Wirtszelle
2 Übersetzen des viralen Erbguts *(RNA)* in DNA
3 Einbau der DNA in das Wirtsgenom
4 Reifung neuer Viruspartikel

Zur Behandlung von HIV-Patienten setzt man gegenwärtig rund 20 Medikamente ein. Grund für die hohe Zahl von Medikamenten, ist die Tatsache, dass Viren immer wieder neue Resistenzen gegen die verabreichten Wirkstoffe entwickeln. Die Ursache dafür sind so genannte Resistenzmutationen – bleibende Veränderungen im viralen Erbgut, die den Krankheitserreger gegen das Medikament schützen. Die Behandlung von HIV-Patienten wird dadurch enorm erschwert. Problematisch ist zudem die Entstehung von Kreuzresistenzen. Dabei entwickelt ein Virus nicht nur gegen das verabreichte Medikament Resistenzen, sondern auch gegen andere Wirkstoffe, die der Patient noch nicht eingenommen hat. In der Abteilung 3 des Max-Planck-Instituts wurden wesentliche Beiträge zur Analyse von HIV-Resistenzen geleistet.

Therapieauswahl

Bei der Wahl eines geeigneten Medikaments müssen Mediziner nicht nur mögliche Nebenwirkungen, sondern vor allem auch Resistenzen des Virus berücksichtigen. Am Max-Planck-Institut für Informatik wurde mit der Software geno2pheno ein Verfahren entwickelt, das auf der Basis des viralen Erbguts die Resistenz des Virus gegen einzelne Medikamente berechnet. Dazu wurde zunächst eine Datenbank aufgebaut, die einander zugeordnete genotypische *(virales Erbgut)* und phänotypische *(Resistenz)* Messdaten enthält. Dann wurden Methoden des maschinellen Lernens verwendet, um die Beziehung zwischen dem viralen Erbgut und der gemessenen Resistenz zu erlernen.

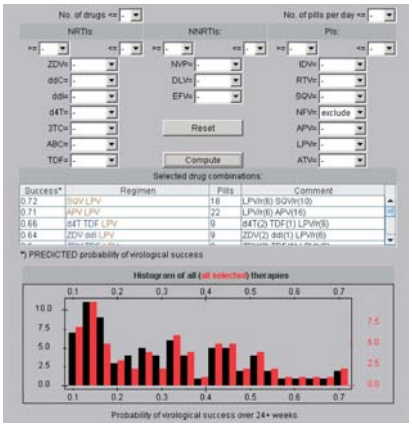


Abbildung 1: Bewertung verschiedener Kombinationstherapien durch die EuResist Prediction Engine

HIV-Medikamente werden heute jedoch nur noch in Kombination verabreicht, damit die Entwicklung von resistenten Virus-Varianten erschwert wird. Um behandelnde Ärzte bei der Auswahl einer wirksamen und lange anhaltenden Kombinationstherapie zu unterstützen, wurde am Max-Planck-Institut für Informatik ein weiteres bioinformatisches Hilfsmittel entwickelt. Auf Grundlage einer Datenbank, die alle relevanten Informationen einer HIV-Therapie speichert, ist ein statistisches Modell erstellt worden, das nach Eingabe des viralen Genoms und der gewählten Medikation die Erfolgswahrscheinlichkeit der Therapie berechnet. Die Software THEO *(THErapy Optimizer)* verwendet dieses Modell und ordnet alle in Betracht kommenden Medikamentenkombinationen nach ihrer Wirksamkeit in eine Rangliste ein [Abbildung 1]. Dabei wird auch berücksichtigt, mit welchen weiteren Mutationen das Virus zukünftig auf die verabreichte Therapie reagiert. Zusätzlich zum detaillierten Wissen über den Patienten erleichtert diese Information dem behandelnden Arzt die Wahl einer optimalen Therapie.

EuResist

Die Vorhersagequalität von statistischen Modellen, wie sie in der Software THEO zum Einsatz kommen, hängt maßgeblich von der Größe der zur Verfügung stehenden Datenbank ab. Daher hat es sich das von der Europäischen Union finanzierte Projekt *EuResist* zum Ziel gesetzt, eine große europaweite Datenbank aufzusetzen, um ausreichend Daten zum Erlernen der Beziehung zwischen viralem Erbgut, Medikamentenkombination und virologischem Ansprechen auf die Therapie zu haben. Zu Beginn des Projektes wurde die EuResist Datenbank durch drei lokale Datenbanken gespeist *(Deutschland, Italien, Schweden)*. Der erfolgreiche Verlauf des Projektes führte dazu, dass sich vier weitere Datenbanken anschlossen, und weitere Datenbanken sollen folgen. Die im Projekt entwickelte Entscheidungshilfe für behandelnde Ärzte besteht aus drei statistischen Modellen. Der Beitrag der Abteilung 3 stellt eine Weiterentwicklung der THEO Software dar. Die Verwendung von drei statistischen Modellen ermöglicht eine stabilere Vorhersage sowie die Abschätzung der Sicherheit der Vorhersage. Analog zu der THEO Software wird nach Übermittlung des viralen Erbguts eine Rangliste der möglichen Therapien erstellt [Abbildung 2]. Hierbei wird sowohl die vorhergesagte Erfolgswahrscheinlichkeit als auch die Schwankung der Vorhersagen berücksichtigt. :::

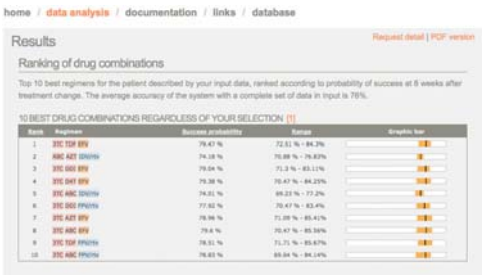


Abbildung 2: Bewertung verschiedener Kombinationstherapien durch die Software THEO

KONTAKT

André Altmann
ABT. 3 Bioinformatik und Angewandte Algorithmik
Telefon +49 681 9325-315
Email altmann@mpi-inf.mpg.de
Internet http://www.euresist.org

Die genetische Grundlage von Krebsleiden

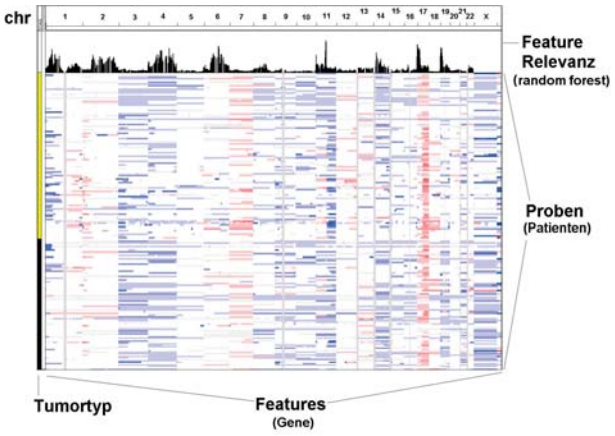
Bei Krebs gerät das Genom durcheinander

Mit Krebs bezeichnet man eine recht diverse Gruppe von Krankheiten, bei denen es zur unkontrollierten Zellvermehrung kommt. Um sich von gesundem Gewebe zu Tumorgewebe zu verändern, müssen eine Reihe von Veränderungen in der Zelle eintreten, die die genetischen Mechanismen der Zellregulation schädigen. Eine bestimmte Sorte von genetischen Veränderungen, die in vielen Krebsgeweben beobachtet wird, ist die Änderung der Kopienanzahl von Genomsegmenten im Zellkern. In der gesunden menschlichen Zelle gibt es genau zwei Kopien aller Chromosome (mit Ausnahme des X- und Y-Chromosoms bei Männern): eine Kopie kommt vom Vater, eine von der Mutter. Krebszellen haben mehr oder weniger Kopien einiger Anteile des Genoms.

Durch **Verlust** von Genomstücken kann die Zelle genetische Elemente einbüßen, die unkontrollierte Zellvermehrung verhindern *(Tumorsuppressorgene)*. Durch **Vergrößerung** der Kopienanzahl von Genen, die die Zellteilung befördern *(Onkogene)*, wird die vermehrte Zellteilung stimuliert. In der Regel müssen eine ganze Reihe verschiedener solcher Veränderungen auftreten, bevor das Gewebe bösartig wird.

Die Entdeckung von Biomarkern

Neuere Entwicklung in der DNA Chiptechnologie sowie in der Sequenzierungstechnologie erlauben die vergleichende genomweite Messung der Kopienanzahl von hunderttausenden Genomregionen in gesunden und krankhaften Zellen. Die Ergebnisse dieser Experimente fördern zutage, welche Genomregionen übermäßige oder mangelhafte Kopienanzahlen aufweisen. Die Muster dieser Veränderungen können vielseitige Informationen über die Art und den Status eines Tumor beherbergen. Solche Muster zu interpretieren ist jedoch sehr schwierig, da der Zusammenhang zwischen dem Veränderungsmuster und der klinischen Ausprägung des Tumors sehr komplex ist. Die biomedizinische Forschung bemüht sich darum, begrenzte und leicht messbare Genomregionen zu identifizieren, die charakteristisch für einen bestimmten Krebstyp sind. Solche



Relevante Abweichungen in Neuroblastoma unter Verwendung von 'random forest' Klassification. Rote und blaue Segmente repräsentieren Amplifikationen und Deletionen. Die lokalen Maxima der Feature-Relevanz zeigen potentielle Biomarker an.

Regionen nennt man *(Krebs-)Biomarker*. Schon heute ergänzen einige Biomarker erfolgreich die klassische Diagnose. Zum Beispiel sind eine Vergrößerung der Kopienanzahl des HER-2 Gens und ein Verlust des p53 Gens Anzeichen für die Entwicklung von Brustkrebs. Die Häufung des MYCN Gens und Veränderung der Kopienanzahl des langen Arms von Chromosom 11 zeigen die Krebsart Neuroblastoma an, usw.

Biomarker zu finden ist schwierig, da eine große Menge experimenteller Daten zu erheben sind und in der Regel nur kleine Anzahlen von Tumorproben zur Verfügung stehen. Hier kommen bioinformatische Methoden zum Einsatz. Am Max-Planck-Institut entwickeln wir statistische Lernmethoden zur Bestimmung und Validierung von genetischen Biomarkern für Krebs.

Die Charakterisierung von Genomveränderungen mit maschinellen Lernmethoden

Maschinelle Lernmethoden zur sogenannten Feature Selektion sind ein ideales Werkzeug für die Suche nach Krebs-Biomarkern. Ein Feature ist ein genetischer Faktor, der experimentell gemessen wird, etwa eine Mutation oder die Löschung einer genomischen Region. Die große Anzahl von Features führt zu einer hohen Dimensionalität der Daten und macht das Problem schwierig. Unsere Lösungen gruppieren die Änderungen in benachbarten Genomregionen und reduzieren dadurch Featureanzahl und Dimension beträchtlich. So konnten wir bestätigen, dass Neuroblastom Tumore sich in jungen und alten Patienten hinsichtlich der genetischen Veränderungen unterscheiden. Für Zellen der Lunge kommen wir zu dem Schluss, dass die Vermehrung des EGFR Gens zur Resistenz gegen bestimmte Krebsmedikamente führt. Unsere Ergebnisse können und müssen durch Laborexperimente bestätigt werden.

In der Kooperation Oncogene mit verschiedenen deutschen Universitäten und Forschungsinstituten, die vom BMBF gefördert wird, suchen wir nach Genen, deren Veränderung mit der Entwicklung von Krebs in Zusammenhang steht und die sich als Ziele medikamentöser Therapien eignen. Wir übernehmen in dem Projekt die Entwicklung statistischer Methoden und ihre Anwendung auf die von anderen Projektpartnern erhobenen experimentellen Daten. :::

KONTAKT

Laura Tolosi
ABT. 3 Bioinformatik und Angewandte Algorithmik
Telefon +49 681 9325-326
Email laura.tolosi@mpi-inf.mpg.de
Internet http://oncogene.bioinf.mpi-inf.mpg.de/

Biomedizin 2.0 – Krebsdiagnostik mit dem Computer

Krebs liegt in den Genen. Aber nicht nur: oft ist es die Verpackung unserer Gene, die Tumore erzeugt. Diese Erkenntnis nutzen Wissenschaftler am Max-Planck-Institut für Informatik, um die Wirksamkeit von Chemotherapie am Computer vorherzusagen.

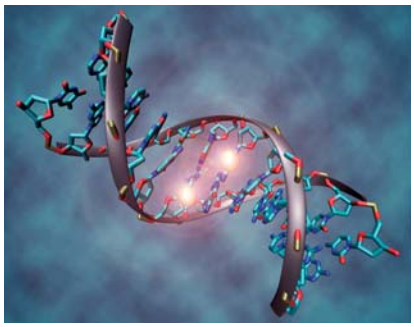
Menschen kooperieren, wenn sie ein gemeinsames Ziel haben — zum Beispiel im Kampf gegen Krebs. Die größte Kooperationsleistung ist aber schon erbracht, bevor wir überhaupt geboren werden: 100 Milliarden Zellen entstehen aus einer einzigen Eizelle und ordnen ihr eigenes Schicksal dem Menschen unter, dessen Teil sie sind. Sie spezialisieren sich, verzichten auf unkontrollierte Vermehrung, und wenn der Befehl zum zellulären Selbstmord kommt, dann werden sie ihn befolgen. Diese selbstlose Zusammenarbeit ist nur möglich, weil alle Zellen eines Körpers dasselbe Erbgut in sich tragen. Damit stehen sie nicht mehr im Sinne Darwins miteinander im Wettbewerb. Denn für das Überleben der gemeinsamen Gene spielt es keine Rolle, ob eine Zelle sich selbst vermehrt oder ihre Schwesterzellen diese Aufgabe übernehmen. Ein komplexer Organismus mit hochspezialisierten Zellen kann so erst entstehen.

Wie aber spezialisieren sich Zellen, wenn sie allesamt dasselbe Erbgut tragen? Die Evolution hat den Weg in die dritte Dimension gewählt: Während die Buchstaben der DNA unverändert bleiben, werden die DNA-Moleküle der Zelle auf vielfältigste Weise neu verpackt, verdrillt und aufgewickelt. Dadurch sind in jeder Zelle nur die jeweils wichtigen Gene zugänglich, während unwichtige oder gar schädliche Gene säuberlich verpackt und archiviert werden. Ein weißes Blutkörperchen kann beispielsweise seine Immun-Gene jederzeit aktivieren, da es sich frei zugänglich mitten im Zellkern befindet. Hingegen ist ihm die Verwendung von gehirnspezifischen Genen strikt untersagt; sie liegen aufgerollt an der inneren Wand des Zellkerns. Schließlich könnte schon die fehlerhafte Aktivierung einzelner Wachstumsgene zur Tumorbildung führen. Umgekehrt ist jedoch auch die versehentliche Deaktivierung wichtiger Gene eine regelmäßige Begleiterscheinung von Krebs.

Grund genug, sich das Verpacken der Gene einmal näher anzuschauen. Aus dieser Fragestellung hat sich in den letzten Jahren das junge Forschungsgebiet der Epigenetik entwickelt – und ein neues Anwendungsfeld für bioinformatische Algorithmen! Um krebsrelevante epigenetische Veränderungen zu finden, muss nämlich zunächst eine Karte der epigenetisch wichtigen Regionen im Erbgut erstellt werden. Für die Kartierung spielen Lernalgorithmen und Mustererkennung eine Rolle, da sie Zusammenhänge in den Daten identifizieren können, wie es sonst nur unser Gehirn vermag. Der Computer vergleicht die Eigenschaften der DNA in epigenetisch zugänglichen Regionen mit denen von epigenetisch verpackten Regionen und lernt eigenständig die wichtigsten Unterschiede.

Mit derartigen Karten vereinfacht sich die Suche nach krebsspezifischen Veränderungen. Entscheidend ist jedoch der nächste Schritt: Bioinformatische Methoden werden auf klinische Daten angewendet, mit dem Ziel, die Krebs-Therapie zu verbessern. In Zusammenarbeit mit Kooperationspartnern vom Bonner Universitätsklinikum ist dies bereits für eine bestimmte Form von Gehirntumoren gelungen. Ein Gen namens MGMT hilft diesen Tumoren dabei, sich gegen Chemotherapie zu verteidigen. Allerdings ist das MGMT-Gen in einigen Tumoren epigenetisch verpackt und deaktiviert, was diese Tumore besonders anfällig für eine aggressive Chemotherapie macht.

Ärzte und Patienten stehen also vor einer schwierigen Frage: Entweder sie entscheiden sich für die stark belastende Chemotherapie, obwohl sie oft keinerlei Wirkung zeigt, oder sie verpassen eine mögliche Behandlungschance. Die bessere Lösung wäre ein Labortest, mit dem sich die Wirksamkeit der Chemotherapie für jeden einzelnen Patienten vorhersagen lässt. Mit Hilfe von bioinformatischen Lernalgorithmen konnte ein derartiger Test konstruiert werden, basierend auf epigenetische Daten für erfolgreich sowie erfolglos behandelte Patienten. Das so abgeleitete Computermodell kann mit großer Genauigkeit vorhersagen, ob ein Tumor sein MGMT-Gen epigenetisch deaktiviert hat und er damit ein leichtes Opfer für eine Chemotherapie ist. Außerdem gelang es, durch statistische Optimierung die Materialkosten der experimentellen Analyse auf wenige Euro pro Patient zu reduzieren. Damit konnte der MGMT-Test am Bonner Universitätsklinikum erstmals in der Diagnose eingesetzt werden und dazu beitragen, dass Patienten individuell optimal behandelt werden. :::



Methylierte DNA-Moleküle sind Träger epigenetischer Information.

Weiterführende Information finden Sie auf folgenden Webseiten:

Epigenetik? (http://epigenome.eu/de/) – Das Wissenschaftsportal des Epigenom-Exzellenznetzwerks der EU

Spurensuche an bösartigen Genen
(http://tinyurl.com/6xq6v2)
Ein Artikel zur epigenetischen Krebsdiagnose, erschienen in der Zeitschrift *MaxPlanckForschung*

KONTAKT

Christoph Bock
ABT. 3 Bioinformatik und Angewandte Algorithmik
Telefon +49 681 9325-325
Email cbock@mpi-inf.mpg.de
Internet http://www.computational-epigenetics.de

Korezeptorvorhersage bei HIV

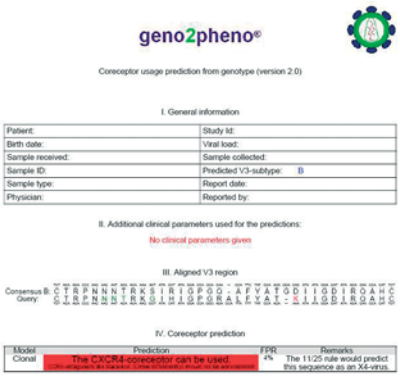
Korezeptorinhibitoren – eine neue Klasse von HIV Medikamenten

Die AIDS Forschung hat in den letzten Jahren eine Fülle neuer Wirkstoffe hervorgebracht. Bis vor kurzem haben diese sich auf zwei Proteine des Virus konzentriert: die Protease und Reverse Transkriptase. Der Nachteil dabei war, dass HIV sehr schnell mutiert und schon einzelne Mutationen innerhalb dieser Proteine zu Resistenzen gegen ein oder mehrere Medikamente führen kann. Mit der Verfügbarkeit von neuen Medikamentenklassen wie z.B. den Integrasehemmern oder Fusionshemmern, die verschiedene Proteine des Virus angreifen, konnte dies ein wenig verbessert werden.

Einem anderen Ansatz folgt die Klasse der sogenannten Korezeptor-Antagonisten, die statt einem viralen ein menschliches Protein blockieren. In diesem Falle ist dies ein sogenannter Korezeptor, ein zelluläres Oberflächenprotein, das HIV zum Eindringen in die Zelle benötigt. Die Hoffnung dabei ist, dass für das Virus zusätzliche Mutationen erforderlich sind, um gegen alle verabreichten Wirkstoffe resistent zu werden.

Viraler Tropismus

Der erste Korezeptor-Antagonist Maraviroc *(Celsentri, Pfizer)* wurde Ende 2007 von den Gesundheitsbehörden der USA und Europa zugelassen. Dieser Wirkstoff bindet spezifisch an den Chemokin-Rezeptor CCR5, der auf menschlichen Zellen exprimiert wird und den bestimmte HIV Stämme, vor allem solche, die in der frühen Infektionsphase dominieren, als Korezeptor verwenden. Neben diesem Rezeptor gibt es allerdings noch einen zweiten Rezeptor *(CXCR4)*, den manche Viren statt CCR5 verwenden können. Je nachdem welchen Korezeptor ein Virus benutzt, wird es als R5- *(CCR5)* oder X4-Virus *(CXCR4)* bezeichnet. Diese Eigenschaft des Virus wird auch als viraler Tropismus bezeichnet. Da X4-Viren den CCR5-Rezeptor nicht brauchen, besitzen sie eine natürliche Resistenz gegen Maraviroc



Vorhersage des Korezeptors mittels geno2pheno[coreceptor]

und daher sollte das Medikament bei diesen auch nicht verabreicht werden. Folglich wird vor Verschreibung von Maraviroc ein sogenannter Tropismus- oder Korezeptortest gemacht, durch den ausgeschlossen werden soll, dass sich X4-Viren im Blut des Patienten befinden.

Tropismustest und geno2pheno[coreceptor]

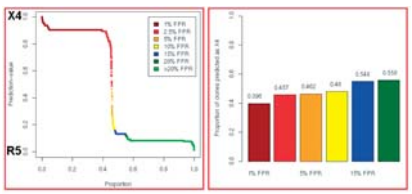
Bis vor kurzem wurde der Tropismus mit Hilfe eines Labortests bestimmt. Wie im Falle der Resistenzbestimmung bei traditionellen Medikamenten gibt es jedoch auch einen computer-basierten Ansatz. Hierbei wird die RNA-Sequenz der den viralen Tropismus bestimmenden Region des viralen Genoms *(die sogenannte V3 Schleife des viralen Hüllproteins gp120, die nur gut 100 genomische Buchstaben umfasst)* im Labor bestimmt *(sequenziert)* und der Tropismus daraus mit dem Computer vorhergesagt. In der Abteilung 3 des Max-Planck-Instituts wurden solche Vorhersagemethoden bereits in der Vergangenheit entwickelt und in das geno2pheno[coreceptor]-Websystem implementiert. Dieser Webservice wurde in den letzten beiden Jahren komplett überarbeitet und um weitere Funktionen ergänzt.

So kann z.B. auch der Immunstatus des Patienten bei der Vorhersage berücksichtigt werden und damit genauer auf den individuellen Fall eingegangen werden.

Die Arbeiten zu geno2pheno[coreceptor] laufen in enger Kooperation mit Virologen aus Deutschland, Europa und Kanada. Unser Webserver ist in den letzten beiden Jahren über 120.000-mal verwendet worden. Diese Popularität spiegelt sich auch in den Deutsch-Österreichischen AIDS Leitlinien wider, in denen es nun als zum Labortest gleichwertiges System für die Tropismustestung empfohlen wird.

Neue Ansätze

In den letzten Jahren haben wir darüber hinaus weitere Ansätze erforscht, um die Vorhersagequalität unseres Servers weiter zu verbessern. Zum einen konnten wir zeigen, dass sich durch die Verwendung eines weiteren Genabschnitts des viralen Hüllproteins, der V2-Schleife, die Vorhersagen signifikant verbessern lassen. Zum anderen haben wir eine völlig neue Sequenziertechnologie verwendet, um tiefer in die Virenpopulation, die sich im einzelnen Patienten befindet, hineinzuschauen. Dieser Ansatz ermöglicht es uns hunderttausende verschiedene Viren gleichzeitig aus einer Blutprobe zu sequenzieren und zu analysieren. Dadurch können kleinste Minoritäten von X4-Viren entdeckt werden, die möglicherweise zum Versagen einer Maraviroc-Therapie führen. :::



Quantitative Analyse des Korezeptorgebrauchs

KONTAKT

Alexander Thielen
ABT. 3 Bioinformatik und Angewandte Algorithmik
Telefon +49 681 9325-308
Email athielen@mpi-inf.mpg.de
Internet http://www.geno2pheno.org

Ursprung und Adaption des neuen Influenza A/H1N1 Virus

Ausbreitung eines neuen Grippe-Virus

Als am 24. April 2009 erstmals in den Nachrichten von einem neuen Grippevirus in Mexiko und in den Vereinigten Staaten berichtet wurde, war noch nicht abzusehen, wie schnell und weitläufig sich das Virus verbreiten würde. Für den zum Subtyp H1N1 gehörende Erreger der so genannten Schweinegrippe werden mittlerweile steigende Infektionszahlen registriert, für Deutschland vermeldete das Robert-Koch-Institut Anfang Herbst 2009 bereits 30.000 registrierte Fälle.

Das Genom des Influenzavirus besteht aus acht einzelnen Segmenten einer einzelsträngigen RNA, welche wiederum elf verschiedene Proteine kodieren [Abbildung 1]. Diese Proteine benötigt der Erreger, um zum Beispiel in die Zellen seines Wirts einzudringen oder sich darin zu replizieren. Der Aufbau der Proteine ist aber auch entscheidend dafür, ob eine Infektion mit dem Virus für den Wirt mild oder sehr problematisch oder gar tödlich verläuft. So war beispielsweise die Spanische Grippe im Jahr 1918 für den Tod von ca. 50 Millionen Menschen verantwortlich, während bei dem aktuellen A/H1N1 Virus bis jetzt meist ein milder Krankheitsverlauf beobachtet wird.

In der Forschungsgruppe Computational Genomics and Epidemiology wurden die Entstehung und die Eigenschaften des neuen Virus anhand seines Genoms unmittelbar nach der Veröffentlichung der ersten RNA-Sequenzen untersucht.

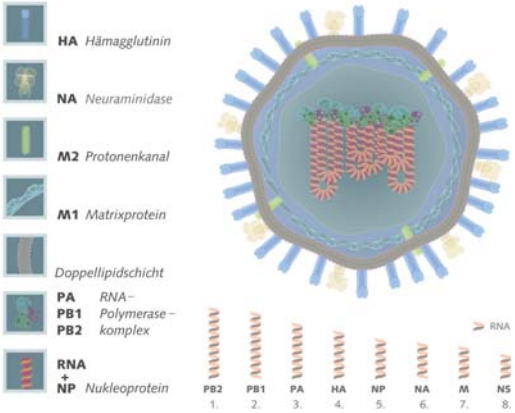
Ursprung des Virus

Influenzaviren befallen nicht nur Menschen, sondern sind auch unter Vögeln und anderen Säugetieren sehr verbreitet. Viren, deren Genom sich an eine bestimmte Wirts-Spezies angepasst hat, springen dabei relativ selten auf eine andere Spezies über. Nach aktuellen Erkenntnissen stammen die acht RNA-Segmente des aktuellen Virus ursprünglich aus einer Vielzahl von Influenza-Stämmen, die unabhängig voneinander in Vögeln, Schweinen und auch im Menschen verbreitet waren und auch teilweise selbst Neuordnungen mehrerer Stämme sind, die seit vielen Jahren in diesen Wirten in unterschiedlichen Teilen der Welt kursieren.

Adaption an den Menschen

Am Max-Planck-Institut für Informatik wurde untersucht, welche Veränderungen im Genom des zuletzt in Schweinen zirkulierenden Virus dazu geführt haben, dass sich der Erreger nun auch von Mensch zu Mensch übertragen lässt. Dazu wurde nach statistisch und biologisch relevanten Unterschieden in den Genomsequenzen von Influenzaviren gesucht, die an den Menschen angepasst sind (wie zum Beispiel die saisonale Grippe oder das 1918 Virus), und solchen, die (noch) nicht adaptiert sind (wie beispielsweise die Vogelgrippe). Solches Wissen ist dringend notwendig, um neue Medikamente entwickeln zu können, die eine weitere Verbreitung des neuen Erregers verhindern könnten, und um in Zukunft Erreger aus anderen Spezies, die in der menschlichen Bevölkerung Fuß fassen könnten, frühzeitig zu identifizieren.

Zusätzlich kursieren zur Zeit aber auch weitaus gefährlichere Influenzaviren, die bis jetzt nur sehr selten auf den Menschen oder gar zwischen Menschen übertragen werden. Sollte sich ein hochgradig pathogenes Virus, wie zum Beispiel die Vogelgrippe, ähnlich an den Menschen anpassen wie das aktuelle H1N1 Virus, könnte das zu einer sehr bedrohlichen Pandemie führen, die ähnliche Dimensionen wie der Ausbruch der Spanischen Grippe vor 90 Jahren hat. Deshalb wird derzeit weiter an Szenarios geforscht, die zu einem solchen Ausbruch führen könnten und nach Mutationen gesucht, die dem Vogelgrippe-Virus eine effiziente Verbreitung in der menschlichen Population ermöglichen könnten. :::



Schematische Darstellung eines Influenza-Viruses

KONTAKT

Alice McHardy
UFG. 1 Informatik für die Genomforschung und Epidemiologie
Telefon +49 681 9325-309
Email mchardy@mpi-inf.mpg.de

Lars Steinbrück
UFG. 1 Informatik für die Genomforschung und Epidemiologie
Telefon +49 681 9325-314
Email lsbrueck@mpi-inf.mpg.de

Christina Tusche
UFG. 1 Informatik für die Genomforschung und Epidemiologie
Telefon +49 681 9325-321
Email ctusche@mpi-inf.mpg.de

Informatik für die Metagenomforschung: Einblicke in die Welt der unkultivierbaren Mikroorganismen

Einleitung

Wussten Sie, dass die überwiegende Mehrheit der Bewohner dieses Planeten mit bloßem Auge nicht erkennbar ist? Es handelt sich hierbei um Mikroorganismen: winzige Lebewesen, die erst mit Hilfe eines Mikroskops sichtbar werden. Man findet sie nahezu überall: Im Eis der Arktis, in heißen Quellen und tiefen Meereskratern, im menschlichen Verdauungstrakt und auf der Haut; ja sogar in der Atmosphäre.

Die Erforschung von Mikroorganismen ist unter agrarwirtschaftlichen, biotechnologischen, umwelttechnischen und medizinischen Aspekten sehr interessant, da diese sich durch eine Vielzahl von einzigartigen Fähigkeiten auszeichnen. Viele solcher Erkenntnisse erhält man bereits durch eine Analyse der Genomsequenz. Dabei gibt es jedoch eine Schwierigkeit: Die meisten Mikroorganismen lassen sich nicht in einer reinen Kultur im Labor anreichern; sie benötigen zum Teil unbekannte Zusatzstoffe, um zu wachsen, oder sind auf die Interaktion mit anderen Organismen in ihrer Umgebung angewiesen. Deshalb ist es notwendig, Zellmaterial und DNA direkt aus deren natürlicher Umgebung zu isolieren. Die genomische Analyse einer Gemeinschaft von Organismen durch Sequenzierung nennt sich Metagenomik. Am Max-Planck-Institut für Informatik werden Verfahren entwickelt, um die bei der Sequenzierung entstehenden genomischen Sequenzfragmente zusammenzusetzen und den verschiedenen Mikroorganismen zuzuordnen.

Die Sequenzierung von mikrobiellen Gemeinschaften

Da nur kürzere Abschnitte eines Genoms direkt sequenziert werden können, wird das isolierte Erbmaterial zuerst in viele kurze Fragmente zerstückelt. Deren Sequenzierung führt zu so genannten *reads*; Sequenzen von weniger als 1000 Nukleotiden Länge. Anhand von wiederholten Abschnitten lassen sich reads zusammenfügen und so längere Bereiche einer Genomsequenz rekonstruieren, was zu einer Vielzahl genomischer Sequenzfragmente von unterschiedlicher Länge führt.

Zuordnung von Sequenzfragmenten

Die Zuordnung der Sequenzfragmente zu den in der Gemeinschaft vorhandenen Organismen wird als *Binning* bezeichnet. Die am Max-Planck-Institut für Informatik entwickelten Verfahren machen sich zum Lösen dieses Problems eine Eigenschaft von genetischen Sequenzen zunutze, die Hinweise auf die Zugehörigkeit eines Fragments gibt: Zählt man das Vorkommen von *Oligomeren* (kurzen Teilwörtern des Genoms mit Längen zwischen 2 und 30 Nukleotiden) in der Genomsequenz eines bestimmten Organismus, so erkennt man, dass manche Oligomere häufiger vorkommen als andere. Mit einem statistischen Lernverfahren namens *Support Vector Machine* (SVM) lässt sich dann ein phylogenetisches Modell von universell vorkommenden Markergenen erstellen, welches die Charakteristika dieser Organismen im Oligomer-Gebrauch beschreibt. Anschließend werden die Fragmente des Metagenoms durch das Modell den ermittelten Klassen zugeordnet.

Anwendungen

Das beschriebene Binning-Verfahren wird zur Analyse vieler Metagenome eingesetzt, zum Beispiel von Bakterien im Darm von Nasutitermes ephratae, einer höheren Termitenart. Dies führte zu der Entdeckung einer Vielzahl von neuen Enzymen, die die Zerlegung von Holz in einzelne Zucker katalysieren, und mit denen sich nun vielleicht effiziente industrielle Verfahren zur Gewinnung von Wasserstoff aus pflanzlichen Materialien entwickeln lassen. Die aktuelle Suche nach industriell einsetzbaren Enzymen wird fortgeführt mit der Analyse mikrobieller Metagenome aus den Mägen von niederen Holzabbauenden Termiten oder des australischen Tammar Wallabys, welches besonders wenig des Treibhausgases Methan beim Verdauen pflanzlicher Materialien als Nebenprodukt produziert.

Ausblick

Neben der Erzeugung von alternativen Treib- und Brennstoffen sind mikrobielle Gemeinschaften auch für viele andere industrielle Prozesse von Interesse: So werden Mikroben zur Abwasseraufbereitung bei der industriellen Plastikproduktion eingesetzt und die Metagenom-Analyse einer mikrobiellen Gemeinschaft in einem neuen und effizienteren Bioreaktor soll das Verständnis der hieran beteiligten biochemischen Prozesse und Organismen vertiefen, um diesen industriellen Prozess weiter zu optimieren. Aktuell werden die existierenden Verfahren weiterentwickelt, um eine Klassifizierung von sehr kurzen Sequenzfragmenten zu ermöglichen. Deren Zuordnung ist aufgrund der wenigen vorhandenen Sequenzinformation besonders schwierig. Weiterhin wird an einem Verfahren gearbeitet, welches eine Zuordnung von neuen, bisher unbekannten Genen in einem Metagenom zu den in der mikrobiellen Gemeinschaft ablaufenden biologischen Prozessen ermöglichen soll. :::

KONTAKT

Alice McHardy
UFG. 1 Informatik für die Genomforschung und Epidemiologie
Telefon +49 681 9325-309
Email mchardy@mpi-inf.mpg.de

Kaustubh Patil
UFG. 1 Informatik für die Genomforschung und Epidemiologie
Telefon +49 681 9325-314
Email patil@mpi-inf.mpg.de
Internet http://cge.mpi-inf.mpg.de/

GARANTIEN

Software soll verlässlich sein. Das wichtigste Kriterium für Verlässlichkeit ist die Korrektheit. Fast genauso wichtig aber ist oft die Performanz: Eine korrekte Antwort, die man nicht rechtzeitig bekommt, ist unnütz. Die Suche nach Korrektheits- und Performanzgarantien gehört für viele Abteilungen des Instituts zu den zentralen Fragestellungen.

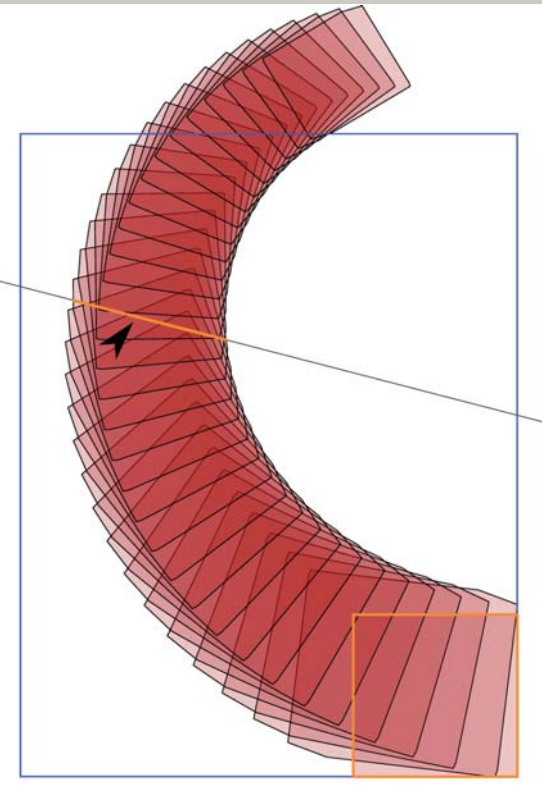
Computer, Netzwerke und mikroprozessorgesteuerte Systeme sind heute ein allgegenwärtiger Teil unseres Lebens. Wir benutzen sie ständig – teils bewusst, wie den Rechner auf dem Schreibtisch, die Internet-Suchmaschine oder das Handy, teils unbewusst, wie die elektronische Steuerung im Auto, im Flugzeug oder in der Waschmaschine. Je mehr wir unser Leben abhängig von Hard- und Software machen, umso mehr stellt sich die Frage, ob das Vertrauen, das wir in diese Produkte setzen, gerechtfertig ist. Können wir garantieren, dass eine Hardware, eine Software oder ein eingebettetes System, das bei Bedarf auch noch mit seiner Umgebung interagiert, wie gewünscht funktioniert? Diese Frage durchzieht die Arbeiten mehrerer Abteilungen des Instituts.

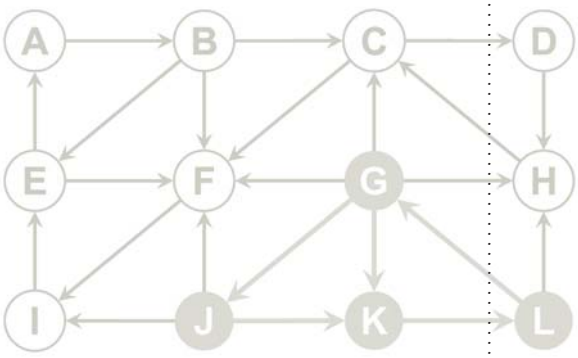
Die einleuchtende Forderung, die man gewöhnlich an eine Hard- oder Software stellt, ist Korrektheit. Wir erwarten, dass die Steuerung eines Flugzeuges korrekt auf die Befehle des Piloten reagiert, dass die Zugsteuerung prinzipiell verhindert dass zwei hintereinander fahrende Züge kollidieren können, oder dass das egoistische Verhalten einer Einzelperson im Internet nicht das ganze Netzwerk lahmlegt. Um solche Eigenschaften nachzuweisen, benötigen wir deduktive Systeme, die überprüfen, ob eine Eigenschaft aus anderen, bereits bekannten Eigenschaften folgt. Ein erster Schritt ist dabei die Modularisierung: die (möglichst automatische) Zerlegung eines großen Problems in kleine, handhabbare Teilprobleme. Ein zweites wichtiges Hilfsmittel ist die Abstraktion. Vereinfacht ausgedrückt versucht man dabei, die unendlich vielen möglichen Daten, mit denen ein Programm arbeiten soll, in endlich viele Klassen geeigneter Ausprägungen zu gruppieren, deren Elemente sich im Wesentlichen gleich verhalten.

Bei Such- und Optimierungsproblemen stellt man fest, dass die Forderungen nach optimalen Ergebnissen und effizienter Berechnung oft nicht gleichzeitig zu erfüllen sind. Manchmal liegt dies an der schieren Größe der zu verarbeitenden Datenmenge, beispielsweise bei der Beantwortung einer komplexen Suchmaschinenanfrage. In anderen Fällen, etwa bei graphentheoretischen Fragestellungen, zeigt sich, dass selbst Probleme von durchaus überschaubarer Größe nicht in akzeptabler Zeit exakt algorithmisch behandelt werden können.

Zusätzliche Komplexität ergibt sich beim Vorliegen einer so genannten Spielesituation, bei der verschiedene Parteien versuchen ihre eigenen, global inkompatiblen Ziele durchzusetzen. Mit diesem neueren Modellierungsansatz lassen sich viele reale Probleme adäquat modellieren und untersuchen. Dazu gehören zum Beispiel das Zusammenspiel der Umwelt mit der sie kontrollierenden Steuerung oder das Verhalten eines Nutzers im Internet im Zusammenspiel mit den Regularien des Netzwerkbetreibers und der anderen Internet Nutzer. In beiden Fällen ist es wichtig trotz divergierender Ziele die Einhaltung globaler Eigenschaften zu garantieren, etwa den fairen Zugang zu den Ressourcen des Internets. :::

BEITRÄGE





Netzwerke entwerfen für egoistische Agenten

Eigennütziges Verhalten in Netzwerksystemen ist ein wichtiges Problem mit großer praktischer Relevanz. Insbesondere stellt sich die Frage, wie das eigennützige Verhalten von Nutzern die Leistung das Systems beeinflusst.

Nehmen wir an, es soll ein Netzwerk für viele Nutzer (Spieler) konstruiert werden, die selbst ihre Kanten auswählen können. Jeder Nutzer hat zwei Knoten, die miteinander verbunden werden sollen. Der Nutzer ist lediglich daran interessiert, seine eigenen Knoten möglichst günstig miteinander zu verbinden. Wenn zwei oder mehr Nutzer eine bestimmte Kante verwenden, teilen sie die Kosten für diese Kante. Es macht also durchaus Sinn für die Nutzer geteilte Kanten zu verwenden. Trotzdem kann es passieren, dass bestimmte Nutzer in einem gegebenen Netzwerk nicht zufrieden sind und lieber „private“ Kanten verwenden würden, die für andere Nutzer nicht im Betracht kommen. Dies kann sogar wiederholt passieren, bis ein Nash-Gleichgewicht erreicht wird, wo alle Nutzer zufrieden sind.

Das „optimale“ Netzwerk wird in vielen Fällen kein Nash-Gleichgewicht sein. Wir können aber versuchen herauszufinden, was das beste Nash-Gleichgewicht ist (also mit den niedrigsten Gesamtkosten für die Konstruktion des Netzwerkes), und dieses Netzwerk als Lösung vorschlagen. Kein Nutzer wird dann von dieser Lösung abweichen wollen, weil alle zufrieden sind.

Die Frage ist jetzt: Wieviel mehr kostet dieses Netzwerk im Vergleich zu dem optimalen Netzwerk im schlimmsten Fall? Das Verhältnis zwischen den Kosten dieser Netzwerken nennt man den Preis der Stabilität *(PdS)*, also den Preis den wir zahlen, damit wir eine stabile Lage erreichen.

Für dieses Problem ist nur in Sonderfällen bekannt, was der PdS ist. Wenn die Kanten gerichtet sind, wächst der PdS logarithmisch in der Anzahl der Nutzer. Auch wenn die Nutzer unterschiedlich wichtig sind, oder alle Nutzer einen Endknoten teilen, ist der PdS nicht konstant. Was passiert aber in dem allgemeinen Fall mit ungerichteten Kanten?

Die Frage, ob der PdS nach oben beschränkt ist oder aber mit der Anzahl von Spielern wächst, bleibt weiterhin offen. Wir können aber zeigen, dass der PdS für dieses Problem wirklich anders ist als für den Fall mit gerichteten Kanten, weil wir für zwei und drei Spieler abweichende Grenzen gezeigt haben. Für zwei Spieler [Abbildung 1] ist der PdS genau 4/3 (gerichtet 3/2), und für drei Spieler ist der PdS höchstens 1.65 (gerichtet 5/3). Außerdem konnten wir eine neue untere Schranke von 1.82 zeigen für den allgemeinen Fall, mit vielen Spielern [Abbildung 2].

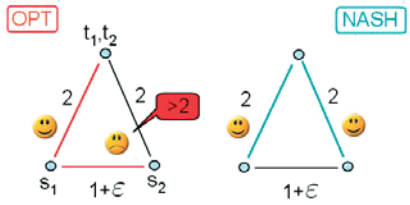


Abbildung 1

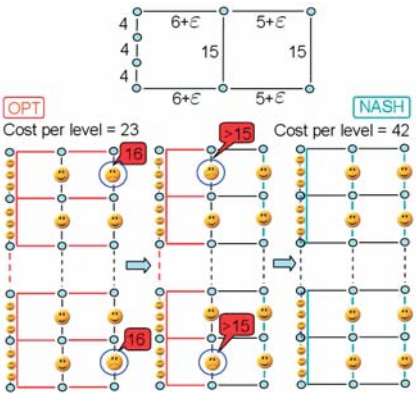


Abbildung 2

Die untere Schranke basiert auf einem Graphen mit *N* Ebenen. Oben ist eine Ebene gezeigt mit den Kosten der einzelnen Kanten. Für jede vertikale Kante gibt es genau einen Spieler, der die Endknoten dieser Kante verbinden möchte. Unten steht links das optimale Netzwerk und rechts das einzige Nash-Gleichgewicht, das es in diesem Netzwerk gibt. :::

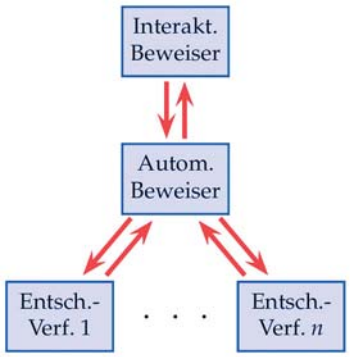
KONTAKT

Rob van Stee
ABT. 1 Algorithmen und Komplexität
Telefon +49 681 9325-105
Email vanstee@mpi-inf.mpg.de

Automatisches Beweisen

Um garantieren zu können, dass eine Hardware oder Software korrekt arbeitet, muss man sie verifizieren – das heißt, die Korrektheit formal nachweisen. Kernelement einer jeden Verifikation ist die Untersuchung, ob bestimmte Eigenschaften aus anderen, bereits bekannten Eigenschaften eines Systems folgen. Mit der Frage, wie man Computerprogramme zum Lösen solcher Beweisaufgaben einsetzen kann, beschäftigen sich Wissenschaftler bereits lange Zeit. Schon seit den fundamentalen theoretischen Ergebnissen von Gödel und Turing zu Beginn des zwanzigsten Jahrhunderts weiß man, dass nicht alles, was im mathematischen Sinne *wahr* ist, auch *beweisbar* ist, und dass nicht alles, was *beweisbar* ist, *automatisch beweisbar* ist. Deduktionssysteme unterscheiden sich dementsprechend deutlich in ihrer Ausdrucksstärke und ihren Eigenschaften: Entscheidungsverfahren sind auf eine bestimmte Art von Daten (etwa reelle Zahlen) spezialisiert und können innerhalb dieses Bereichs garantiert die Korrektheit oder Inkorrektheit einer Aussage nachweisen. Automatische Beweiser für die so genannte erststufige Logik können mit beliebigen, in einem Programm definierten Datentypen umgehen. Hier steht aber nur fest, dass sie einen Beweis finden, falls er existiert; falls keiner existiert, dann suchen sie möglicherweise erfolglos weiter, ohne jemals anzuhalten. Interaktive Beweiser arbeiten mit noch ausdrucksstärkeren als der erststufigen Logik. Sie funktionieren allerdings nur mit Benutzerunterstützung und ohne jede Vollständigkeitsgarantie.

Es ist offensichtlich, dass für praktische Anwendungen im allgemeinen die Kombination aller Methoden nötig ist: Man braucht interaktive Beweiser, um Probleme komfortabel in einer ausdrucksstarken Logik formulieren zu können und um größere Beweise durch eine Auswahl geeigneter Strategien steuern zu können. Man benötigt automatische Beweiser, um den Automatisierungsgrad des interaktiven Beweisers so weit wie möglich zu erhöhen. Und schließlich benötigt man Entscheidungsverfahren, um beispielsweise rein arithmetische Teilaufgaben effizient zu lösen. In unserer Arbeitsgruppe versuchen wir die logischen sowie technischen Schwierigkeiten zu überwinden, die eine solche Kombination mit sich bringt.



Kombination deduktiver Systeme

Neben der Kombination von Deduktionssystemen beschäftigen wir uns zudem seit langem intensiv mit dem automatischen Beweisen. Wie arbeitet ein automatischer Theorembeweiser? Ein Programm zu schreiben, das aus gegebenen Formeln neue Formeln logisch korrekt ableitet, ist nicht schwer. Eine logisch korrekte Ableitung ist allerdings nicht unbedingt eine sinnvolle Ableitung. Wer zum Beispiel $2 \cdot a + 3 \cdot a$ erst in $2 \cdot a + 3 \cdot a + 0$ und dann in $2 \cdot a + 3 \cdot a + 0 + 0$ umwandelt, macht zwar keinen Rechenfehler, kommt seinem Ziel aber keinen Schritt näher. Die eigentliche Herausforderung besteht also darin, aus unendlich vielen *korrekten* Ableitungen die wenigen *sinnvollen* Ableitungen herauszusuchen. Dabei stellt man zunächst fest, dass es nützlich ist, Gleichungen so anzuwenden, dass sich das Ergebnis vereinfacht, also etwa „$x + 0 = x$“ nur von links nach rechts und nicht umgekehrt.

$$x + 0 = x$$
$$x + (-x) = 0$$
$$\frac{x \cdot z}{y \cdot z} = \frac{x}{y}$$

kompliziert → einfach

Gleichungsanwendung

Dieser Ansatz reicht allerdings nicht immer aus. Deutlich wird das beispielsweise bei der Bruchrechnung: Bekanntlich muss man einen Bruch hin und wieder erweitern, bevor man damit weiterrechnen kann. Beim Erweitern passiert aber genau das, was man eigentlich vermeiden möchte: Die Gleichung $(x \cdot z) / (y \cdot z) = x / y$ wird von rechts nach links angewendet – aus einem einfachen Ausdruck wird ein komplizierterer. Der 1990 von Bachmair und Ganzinger entwickelte Superpositionskalkül bietet einen Ausweg aus diesem Dilemma. Einerseits rechnet er vorwärts, andererseits aber identifiziert und repariert er systematisch die möglichen Problemfälle in einer Formelmenge, für die ein Rückwärtsrechnen unvermeidbar sein könnte. Superposition ist damit die Grundlage fast aller heutigen Theorembeweiser für erststufige Logik mit Gleichheit. Das gilt auch für unsere am Institut entwickelten Beweiser SPASS und Waldmeister. Derzeit beschäftigen wir uns nicht nur mit der oben angesprochenen Kombinationsproblematik, sondern insbesondere auch mit speziellen Optimierungstechniken für verschiedene Anwendungen, wie beispielsweise die Analyse von Netzwerkprotokollen oder das Rechnen in großen Ontologien (siehe auch *„Entscheidungsverfahren für Ontologien“*, Seite 60). :::

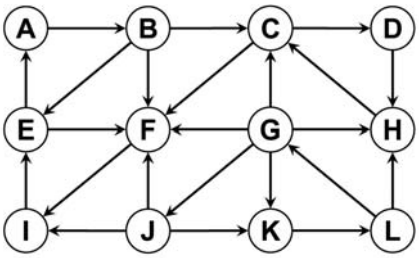
KONTAKT

Uwe Waldmann
FG. 1 Automatisierung der Logik
Telefon +49 681 9325-205
Email uwe@mpi-inf.mpg.de

Model Checking für hybride Systeme

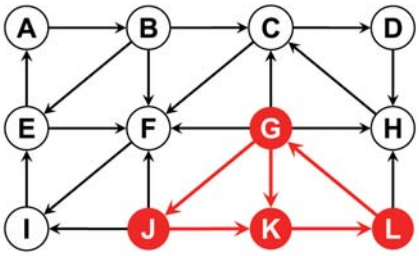
Das Verhalten vieler technischer Systeme lässt sich beschreiben, indem man ihre Zustände und ihre Zustandsübergänge angibt. Eine Waschmaschine kann sich beispielsweise im Anfangszustand (Tür geöffnet, Trommel leer, Maschine ausgeschaltet) befinden, oder auch im Zustand „Waschgang“. Zustandsübergänge sind Wechsel von einem Zustand in einen anderen; diese können automatisch stattfinden (z.B. der Wechsel Waschgang/Abpumpen) oder durch äußere Einwirkung (z.B. weil ein Schalter betätigt wird). Nun gibt es unter den denkbaren Zuständen auch solche, die offenbar unerwünscht sind, z.B. „Wasserzulauf geöffnet, Tür geöffnet“. Wenn wir nachweisen wollen, dass das Gerät *sicher* ist, dann müssen wir zeigen, dass es unmöglich ist, aus dem Anfangszustand durch irgendwelche Zustandsübergänge in einen derartigen *unsicheren* Zustand zu geraten.

Wie kann man dabei vorgehen? Wir können Zustände und Zustandsübergänge durch einen *Graphen* beschreiben: Wir symbolisieren jeden Zustand durch einen Kringel und jeden möglichen Zustandsübergang durch einen Pfeil. Das Ergebnis sieht vielleicht wie folgt aus:



Beispiel

Angenommen, A ist der Anfangszustand und L ist unsicher. Was sind die Vorgängerzustände von L, also die Zustände, von denen aus ein Pfeil nach L zeigt? Offenbar gibt es nur einen solchen Zustand, nämlich K. Also ist auch K unsicher, denn sollte K erreichbar sein, dann können wir von dort aus L erreichen. Ebenso sind die Vorgängerzustände von K, nämlich G und J unsicher. Wenn wir nun die Vorgängerzustände von G und J (also L und G) anschauen, dann stellen wir fest, dass diese bereits als unsicher bekannt sind. Das bedeutet, es gibt keine Möglichkeit, von außen in die Zustandsmenge $\{G,J,K,L\}$ hineinzukommen.



Unsichere Zustände

Damit ist klar, dass $\{G,J,K,L\}$ die einzigen Zustände sind, die per se unsicher sind, oder von denen aus man einen unsicheren Zustand erreichen kann. Alle anderen Zustände, insbesondere auch der Anfangszustand A, sind sicher. Man bezeichnet dieses Verfahren, die Sicherheit eines Systems nachzuweisen, als *Model Checking* oder *Modellüberprüfung*.

Leider wird in der Praxis die Anzahl der Zustände schnell zum Problem: Nehmen wir an, dass das Verhalten einer elektronischen Steuerung von 60 Bits $(b_1,...,b_{60})$ abhängt, von denen jedes den Wert wahr oder falsch haben kann. Nun sind 60 Bits nicht viel, aber wenn man alle möglichen Kombinationen daraus betrachtet, kommt man zu 2^{60} Zuständen; das sind über eine Trillion. Das *symbolische Model Checking* bietet hier einen Ausweg: Wir zählen die Mengen der unsicheren Zustände nicht wie oben explizit auf, sondern wir repräsentieren sie stattdessen *symbolisch* durch logische Formeln. Beispielsweise steht die Formel

b_5 und nicht b_{32}

für alle Zustände, bei denen b_5 wahr und b_{32} falsch ist (das sind eine Vierteltrillion Zustände, die man anderenfalls explizit aufzählen müsste!). Auch die Zustandsübergänge kann man symbolisch berechnen: Wenn wir etwa wissen, dass b_5 auf wahr gesetzt werden darf, falls b_3 oder b_4 wahr ist, dann ist auch jeder Zustand unsicher, der die Formel

(b_3 oder b_4) und nicht b_{32}

erfüllt. Wir erhalten also die Vorgängerzustände der obigen Zustandsmenge einfach, indem wir b_5 durch (b_3 oder b_4) ersetzen.

Kann man auf diese Weise auch die Sicherheit einer Fahrzeugsteuerung nachweisen? Eine solche Steuerung ist ein *hybrides* System: wir haben es nicht nur mit *diskreten* Zustandswechseln (Beschleunigen/Bremsen) zu tun, sondern auch mit Variablen wie der Geschwindigkeit, die sich kontinuierlich ändern können. In den Zustandsformeln müssen also auch numerische Variablen vorkommen, z.B.

b_5 und nicht b_{32} und $(x_2 > 50)$

Die diskreten Zustandsübergänge können wir mit solchen Formeln wie bisher behandeln. Wie sieht es mit den kontinuierlichen Zustandsübergängen aus? Solange das dynamische Verhalten mathematisch gesehen einfach sind (zum Beispiel, wenn sich die Geschwindigkeit gleichförmig ändert), kann man hier ein Verfahren einsetzen, das als *Quantorenelimination* bekannt ist; man braucht allerdings einige technische Tricks, um zu verhindern, dass man zu schnell zu große Formeln erhält. In unserer Arbeitsgruppe beschäftigen wir uns zur Zeit insbesondere mit der Entwicklung von Verfahren, die auch bei einem komplizierteren dynamischen Verhalten anwendbar sind. :::

KONTAKT

Uwe Waldmann
FG. 1 Automatisierung der Logik
Telefon +49 681 9325-205
Email uwe@mpi-inf.mpg.de

Modulares Beweisen in komplexen Theorien

Die großen Fortschritte in der Entwicklung der Informationstechnik haben dazu geführt, dass heutzutage komplexe, rechnergesteuerte Systeme fast überall eingesetzt werden: im Haushalt, in Autos, Zügen, Flugzeugen oder Kraftwerken. Insbesondere in den letztgenannten sicherheitskritischen Bereichen können Fehler katastrophale Folgen haben. Es ist deshalb sehr wichtig, das korrekte Funktionieren solcher Systeme zu garantieren, das heißt mathematisch zu beweisen. Eigentlich wäre es wünschenswert, solche Korrektheitsbeweise völlig automatisch vom Rechner durchführen zu lassen. Fundamentale theoretische Ergebnisse von Gödel, Church und Turing zeigen aber, dass das nicht möglich ist. Für konkrete Anwendungsbereiche existieren jedoch effektive automatische Verifikationsverfahren.

Unser Ziel ist es, Rahmenbedingungen zu identifizieren, unter denen effiziente Verifikationsverfahren für komplexe Systeme existieren. Die formale Beschreibung eines komplexen Systems ist aus Teilen zusammengesetzt, die verschiedenen Bereichen entstammen; so finden sich beispielsweise numerische Formeln neben Aussagen über Datenstrukturen. Es ist daher sehr wichtig, effizient in komplexen Theorien, die als Kombinationen verschiedener Bestandteile entstehen, schlussfolgern zu können. Wir sind daran interessiert, Beweisverfahren zu entwickeln, die die modulare Struktur der komplexen Theorien ausnutzen, und es erlauben, spezialisierte Beweiser für das Schlussfolgern in den Teiltheorien zu benutzen. Solche modularen Verfahren sind besonders flexibel und effizient und in vielen Bereichen anwendbar (wie etwa in der Mathematik, Verifikation oder Wissensrepräsentation [Abbildung1].

Die einfachste Form der von uns betrachteten komplexen Theorien sind Erweiterungen einer gegebenen Theorie (hier als *Basistheorie* bezeichnet) mit zusätzlichen Funktionen. Theorieerweiterungen kommen zum Beispiel in parametrischen Ansätzen zur Verifikation reaktiver oder hybrider Systeme vor, in denen

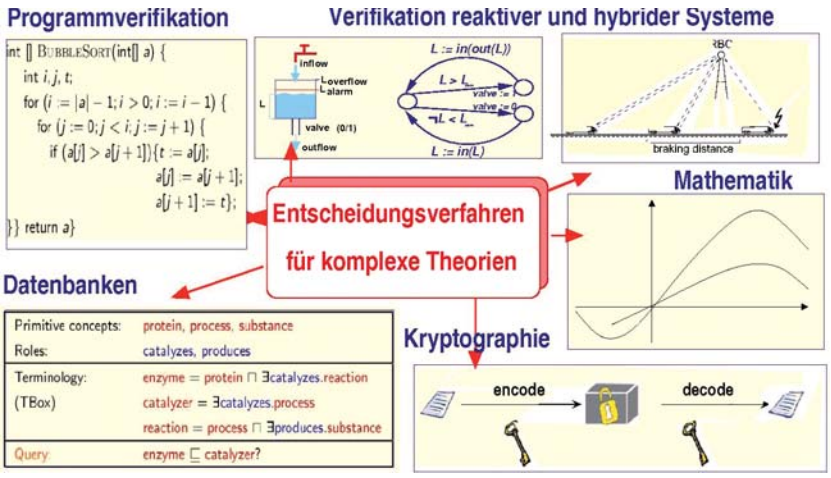


Abbildung 1: Anwendungsbereiche

bestimmte Größen (Zeit, Geschwindigkeit) sowie ihre Veränderungen als Parameter betrachtet werden. Alternativ kann auch die Anzahl von Komponenten ein Parameter sein. Im Allgemeinen ist es schwierig, solche Erweiterungen zu behandeln, auch wenn effiziente Verfahren für die Basistheorie vorhanden sind. Wir gehen die Lösung dieses Problems an, indem wir eine Klasse von Theorieerweiterungen identifizieren, in denen es möglich ist, das ursprüngliche Problem auf ein Problem im Basisbereich zu reduzieren. Dieses kann dann mit einem für die Basistheorie spezialisierten Verfahren gelöst werden. Das allgemeine Prinzip eines solchen hierarchischen Verfahrens ist in Abbildung 2 dargestellt.

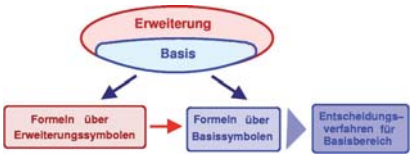


Abbildung 2: Hierarchisches Schließen

Darüber hinaus entwickeln wir Verfahren für modulares Schließen in Kombinationen von Theorien. Wenn wir zum Beispiel bei der Programmverifikation eine Aussage beweisen wollen, in der sowohl von Listen von Zahlen als auch von Arrays von Zahlen die Rede ist, dann können wir dieses Problem in zwei Teilprobleme zerlegen – eines mit Listen und Zahlen, eines mit Arrays und Zahlen – und für jedes davon ein existierendes Beweisverfahren benutzen. Wenn beide Verfahren nun ausreichend viele Informationen über die gemeinsamen Daten (also die Zahlen) austauschen, dann ist gewährleistet, dass zum Schluss die Einzellösungen beider Verfahren zu einer Gesamtlösung kombiniert werden können. Diese Idee ist in Abbildung 3 dargestellt.

Abbildung 3: Modulares Schließen

Unsere theoretischen Beiträge bilden die Basis für die Entwicklung von praktisch einsetzbaren Verifikationswerkzeugen für die Verifikation sicherheitskritischer Systeme, insbesondere im Rahmen des SFB Transregio Projektes AVACS *(Automatic Verification and Analysis of Complex Systems)*. :::

KONTAKT

Viorica Sofronie-Stokkermans
FG. 1 Automatisierung der Logik
Telefon +49 681 9325-207
Email sofronie@mpi-inf.mpg.de

GEOMETRIE

Die Verarbeitung geometrischer Daten ist ein klassisches Gebiet in der Informatik. Gerade dadurch ergibt sich eine große Bandbreite an Forschungsfragen. Aktuelle Forschungsprobleme reichen von der exakten Beschreibung geometrischer Objekte bis hin zu statistischen Methoden zur Interpretation unsicherer geometrischer Informationen.

Das Max-Planck-Institut für Informatik beschäftigt sich seit seinem Bestehen mit verschiedenen Fragen der geometrischen Datenverarbeitung. Grundsätzlich geht es dabei um die Verarbeitung von Datensätzen, die geometrische Objekte im Raum beschreiben. Geometrische Datensätze können dabei z.B. Linien in einer zweidimensionalen Zeichnung sein, dreidimensionale Körper in einer virtuellen Welt oder auch hochdimensionale Stichprobenpunkte aus einer statistischen Datenanalyse. Die Probleme bei der Verarbeitung diese Daten reichen von algorithmischen Grundfragen, wie der effizienten und exakten Repräsentation von geometrischen Objekten aller Art, über die manuelle Modellierung von Objekten bis hin zur automatischen Rekonstruktion mit Techniken aus der Mustererkennung.

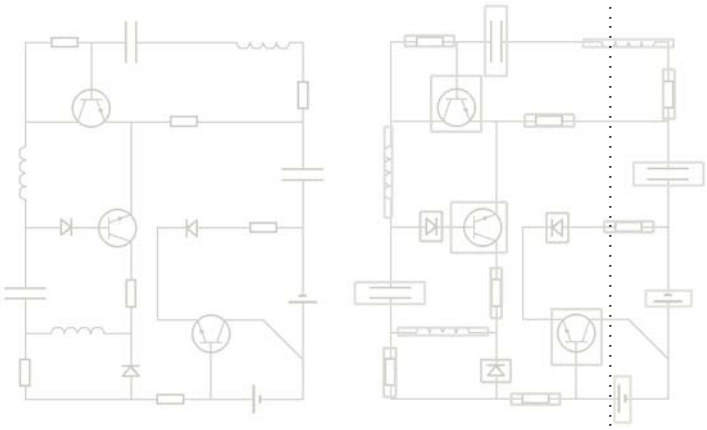
Im Folgenden werden einige der aktuellen Forschungsarbeiten exemplarisch vorgestellt: Ein wichtiges Gebiet ist die Repräsentation geometrischer Objekte. So ist es z.B. mit industriell verfügbarer CAD Software bis heute möglich, Schnitte von geometrischen Objekten exakt und zuverlässig zu berechnen. Dies mag auf den ersten Blick erstaunen, aber die Probleme liegen hier in den Genauigkeitsanforderungen. Traditionelle numerische Algorithmen haben Rundungsfehler die bei manchen Eingabedaten zu intolerablen Fehlern führen. Und leider sind diese „ungünstigen Eingaben“ in der Praxis gar nicht so selten, wie man hofft. In der hier beschriebenen Forschung wurden neue Algorithmen und Datenstrukturen entwickelt, die die Qualität der Ausgabe garantieren können.

Ein geradezu komplementäres Gebiet ist die geometrische Mustererkennung: Hier gibt es keine exakten Eingabedaten sondern oft im Gegenteil sehr unsichere Daten, und ein Algorithmus muss verschiedene unsichere Informationen zu einem Gesamtbild zusammenfassen. Im Folgenden werden mehrere Beispiele aus diesem Problemfeld gezeigt: So sollen in der bildbasierten 3D-Szenenanalyse dreidimensionale Modelle allein aus Fotos oder Videosequenzen rekonstruiert werden. Die Interpretation eines einzelnen Bildes ist mehrdeutig, da der Abstand zu den beobachteten Punkten des Bildes in einem Foto natürlich nicht bekannt ist. Indem Vorwissen eingebracht wird (was kann man von der zu rekonstruierenden Geometrie a priori erwarten?), und ggf. mehrere verschiedene Ansichten kombiniert werden, ergibt sich eine klarere Rekonstruktion. Ähnliche Techniken werden auch für die Erkennung von Gesichtern benutzt: Zunächst wird eine große Datenbank an Gesichtsformen (als 3D-Scans) angelegt. Mit diesem Vorwissen über den „Raum typischer Gesichter“ können Erkennungs- und Rekonstruktionsaufgaben wesentlich besser gelöst werden.

Ein ebenfalls verwandtes Problem ist die Untersuchung von 3D-Geometrie auf Symmetrie: Hier ist die Aufgabe, in einem 3D Modell „Bauteile“ zu finden, die mehrmals in ähnlichen Varianten vorkommen, ohne das die Form dieser Teile dem Algorithmus a priori bekannt ist. Wiederum wird aus potentiell fehlerbehafteten Daten und zusätzlichem Vorwissen (z.B. der Annahme, welche Deformationen der symmetrischen Teile auftreten können) eine Lösung berechnet.

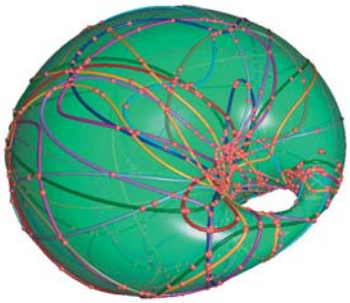
Diese Auswahl an Beträgen zur aktuellen Forschung soll zeigen, dass das traditionelle Gebiet der geometrischen Datenverarbeitung auch heute noch viele Herausforderungen stellt, und dass Lösungen relevant für breite Teile der modernen Informatik sind. :::

BEITRÄGE



Effiziente und exakte Algorithmen für Kurven und Flächen

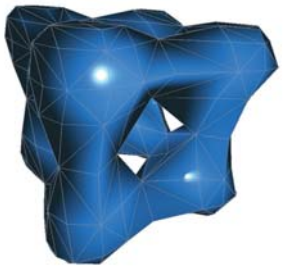
EXACUS (*Efficient and Exact Algorithms for Curves and Surfaces*) und CGAL (*Computational Geometry Algorithms Library*) bezeichnen umfangreiche Sammlungen von C++ Software Bibliotheken zur Untersuchung von Kurven und Flächen. Beide Projekte haben ihren Ursprung in früheren EU Projekten (CGAL, ECG, ACS). EXACUS wurde im Rahmen von ECG am Max-Planck-Institut für Informatik gegründet und dient seit nunmehr fast einem Jahrzehnt als Entwicklerplattform. In den letzten Jahren wurden die wesentlichen Funktionalitäten von EXACUS nach CGAL übertragen, welches inzwischen die grundlegende Entwicklerumgebung für eine Vielzahl von führenden europäischen Forschungsgruppen (INRIA, Berlin, Tel Aviv, Groningen, Zürich, Athen und MPII) in diesem Arbeitsbereich darstellt. Insbesondere im Hinblick auf die Kooperation mit unseren EU Partnern ist auf diesem Wege eine Basis für zukünftige gemeinsame, erfolgreiche und effektive Softwareprojekte gegeben.



Arrangement auf einem Dupin-Cyclide, erzeugt durch den Schnitt mit algebraischen Flächen

In der klassischen algorithmischen Geometrie liegt der Fokus auf dem Studium von linearen Objekten, wie Geraden und Ebenen. Dieser Ansatz wird nun auf die Betrachtung gekrümmter Objekte ausgedehnt. Bei der Implementierung geometrischer Algorithmen treten häufig Schwierigkeiten auf, die sich aus Rechenungenauigkeiten, als Folge von Rundungsfehlern, ergeben: Falsche Resultate, Abstürze oder nicht-terminiernde Programme können die Folge sein. Während sich bei der Untersuchung linearer Objekte in vielen Fällen die Algorithmen unter Verwendung exakter Numerik implementieren lassen, stößt man im Falle von gekrümmten Objekten schnell an die (durch die Mathematik vorgegebenen) Grenzen. Ein Grund ist, dass Lösungen von allgemeinen algebraischen Gleichungssystemen nicht mehr durch Wurzelterme repräsentiert werden können und ein weiterer, dass Berechnungen innerhalb allgemeiner Körpererweiterungen nicht so einfach effizient umzusetzen sind. In bisherigen Ansätzen, so auch bei allen kommerziellen CAD Systemen, setzt man ausschließlich auf numerische Algorithmen, die nur dann korrekte Ergebnisse erzielen, wenn man einen „hinreichend gutartigen" Input voraussetzt. Beispiele zeigen jedoch, dass es durchaus seine praktische Berechtigung hat, eine solche Voraussetzung nicht als gegeben anzusehen. In unserer Arbeit kombinieren wir symbolische Verfahren aus der Computeralgebra mit garantierten numerischen Verfahren und effizienten Algorithmen aus der algorithmischen Geometrie. Auf diesem Wege ist es möglich, alle Fälle korrekt und meistens auch schnell zu untersuchen.

Zum jetzigen Zeitpunkt können wir die Topologie einer ebenen algebraischen Kurve und einer Fläche beliebigen Grades exakt bestimmen. Kurven werden hierbei in Teilsegmente zerlegt, welche charakteristische Punkte, wie Selbstschnitte oder Extremstellen, verbinden. Für mehrere Kurven liefert der Algorithmus eine so genannte Arrangementberechnung, d.h. eine topologische Beschreibung der Zerlegung der Ebene durch die gegebenen Kurven unter Angabe der Beziehung zwischen einzelnen Teilstücken. Darüber hinaus verfügen wir über eine exakte Visualisierung, die wir mit Hilfe eines Webinterfaces der Allgemeinheit zur Verfügung stellen. Die erzielten Ergebnisse in der Ebene lassen sich auch auf parametrisierbare Flächen übertragen. Genauer heißt das, dass wir Arrangements berechnen können, die durch den Schnitt beliebiger Flächen mit beispielsweise einer Kugel oder eines Torus erzeugt



Triangulierung einer algebraischen Fläche

werden. Diese Art von Arrangementberechnungen spielen eine zentrale Rolle in CAD Systemen, bei denen Boolesche Operationen auf Polygonen mit gekrümmten Rändern die grundlegenden Operationen darstellen.

Darüber hinaus beschäftigen wir uns mit dem Studium von algebraischen Flächen. Unsere Arbeitsgruppe verfügt über die weltweit einzige Implementierung, die eine exakte Triangulierung einer beliebigen Fläche bereitstellt. Ein Framework zur Arrangementberechung von Flächen sowie Implementierungen speziell für Quadriken, d.h. Flächen vom Grad 2, sind außerdem verfügbar. Wie bereits angedeutet, stellen die Berechnungen in algebraischen Erweiterungskörpern die größten Hürden hinsichtlich der Effizienz solcher Systeme dar. Ein Herausstellungsmerkmal unserer Forschungsgruppe ist die konsequente Verwendung approximativer (aber garantierter) Verfahren, welche rechenintensive symbolische Methoden ersetzen. Zu solchen approximativen Verfahren gehören beispielsweise Bitstream Solver. Diese ermöglichen die Bestimmung der Nullstellen eines Polynoms, dessen Koeffizienten nur (wenn auch beliebig genau) approximiert werden können. In Zukunft wollen wir vermehrt solche Verfahren entwickeln und mit robusten numerischen Methoden kombinieren. Dabei sollen numerische Verfahren den Hauptteil der Arbeit übernehmen und die exakten Methoden nur der Zertifizierung der Ergebnisse dienen. Erste vielversprechende theoretische Ansätze sind hierbei bereits entwickelt und warten auf ihre Umsetzung. :::

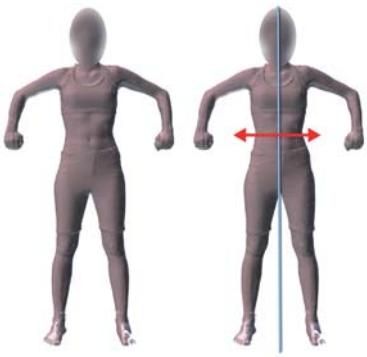
KONTAKT

Michael Sagraloff
ABT. 1 Algorithmen und Komplexität
Telefon +49 681 9325-106
Email msagraloff@mpi-inf.mpg.de
Internet http://www.mpi-inf.mpg.de/EXACUS

Partielle Symmetrien in deformierbaren Objekten

Symmetrien

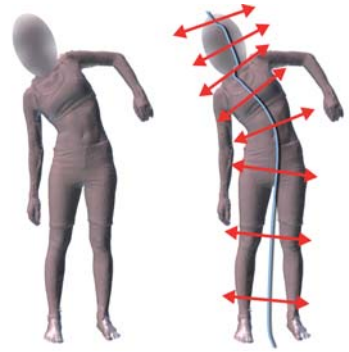
Symmetrien sind fester Bestandteil unserer Welt. Sie sind fest verankert in der menschlichen Wahrnehmung und Denkweise. Sie jedoch maschinell zu erkennen, erweist sich als besonders anspruchsvoll. Dabei würden einige Applikationen davon profitieren können: Objekterkennung und Verbesserung von Scans (Löcher füllen und Qualitätsverbesserung). Sowohl in der Computergrafik als auch in der Computer Vision wurden extrinsische Symmetrien sehr extensiv untersucht. Extrinsische Symmetrien sind abhängig von der aktuellen Pose [Abbildung], und sind im euklidischen Raum definiert: Translationen, Rotationen, Spiegelungen und Skalierungen, also affine Transformationen. Intrinsische Symmetrien „leben" dagegen auf Mannigfaltigkeiten. Wir untersuchen zunächst Symmetrien an deformierbaren Objekten bei denen die Länge zwischen zwei Punkten durch die Deformation erhalten bleibt. In dieser Klasse gehören näherungsweise z.B. artikulierte Objekte. Die zu untersuchenden Daten sind unstrukturierte Punktwolken wie sie z.B. von Rangescannern geliefert werden.



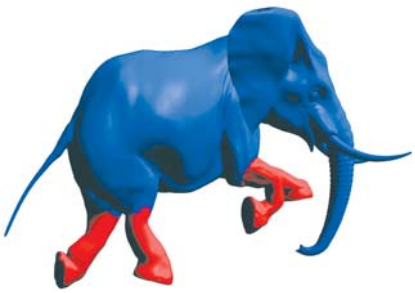
Dreidimensionale Scans, die extrinsische Symmetrien darstellen. Die Symmetrieebene ist durch eine Linie angedeutet.

Graphenmodell

Die Punktwolke (die das Objekt repräsentiert) wird gleichmäßig, spärlich abgetastet, um darauf ein Netzwerk auf der Oberfläche des Objektes zu bilden. Auf diesem Netzwerk wird ein Markov-Zufallsfeld (MRF) definiert, welches höchstens Interaktionen zwischen zwei Punkten modelliert. Die Anzahl der Zustände, die jeder Knoten annehmen kann, wird durch ein gleichmäßiges, aber dieses Mal dichtes Abtasten desselben Objektes bestimmt. Markov-Zufallsfelder liefern eine bedingte Wahrscheinlichkeitsverteilung, wobei die Wahrscheinlichkeit eines Zufallsfeldes (3D Punkt im Netzwerk) von seiner direkten Nachbarschaft abhängt. Eine a priori Wahrscheinlichkeit wird für die Punkte benutzt, um die örtlichen Beziehungen innerhalb der Punktwolke zu modellieren. In unserem Fall sind es längenerhaltende Priors. Die Likelihood wird durch lokale Deskriptoren an den 3D Punkten bestimmt. Die a posteriori Wahrscheinlichkeit über alle intrinsischen Abbildungen des Objektes auf sich selbst beschreibt die Symmetriestruktur des Objektes: die Maxima der Verteilung entsprechen möglichen Symmetrien. Hiermit erhalten wir eine probabilistische Formalisierung für partielle Symmetrien in deformierbaren Objekten. Aus dem formalen Modell lässt sich direkt ein Algorithmus ableiten, der Symmetrien dadurch findet, dass er eine approximative Repräsentation solcher Maxima berechnet. :::



Dreidimensionale Scans, die intrinsische Symmetrien darstellen. Die Hände sind immer noch zueinander symmetrisch, obwohl sie deformiert sind. Es gibt jedoch keine Symmetrieebene im extrinsischen Sinn mehr.



Extrahierte intrinsische Symmetrien bei einem Elefantenmodell

KONTAKT

Ruxandra Lasowski
ABT. 4 Computergrafik
Telefon +49 681 9325-406
Email lasowski@mpi-inf.mpg.de

Michael Wand
ABT. 4 Computergrafik
Telefon +49 681 9325-408
Email mwand@mpi-inf.mpg.de

Bildbasierte 3D-Szenenanalyse

Im Rahmen dieses Forschungsschwerpunktes werden neuartige Methoden und Werkzeuge zur bildbasierten 3D-Szenenanalyse entwickelt. Dabei umfasst die Szenenanalyse erstens die Schätzung der Bewegung der Kamera(s), zweitens die Schätzung der statischen Szenengeometrie und der Bewegungs- und Formschätzungen einzelner bewegter Objekte sowie drittens die Schätzung der Beleuchtung. Solche Werkzeuge und Algorithmen können zum Beispiel bei der Erzeugung von Spezialeffekten bei der Film- und Fernsehproduktion verwendet werden. In spezialisierter Form können solche Algorithmen aber auch bei Land-, Wasser- und Luftfahrzeugen oder Robotern im häuslichen, industriellen und medizinischen Bereich Anwendung finden.

Interaktive 3D-Rekonstruktion mit Hilfe von orthographischen Ansichten

Ziel dieses Projektes ist, eine aufgenommene Bildfolge oder ein Video von einem Objekt zu nutzen, um die 3D-Modellierung des Objektes zu vereinfachen. Dazu werden automatisch aus einer Bildfolge orthographische Ansichten des Objektes erstellt (z.B. Vorderansicht, Seitenansicht, usw.). Diese orthographischen Ansichten können leicht in beliebige Modellierungs- und Animationspakete importiert werden und erlauben dem Benutzer, mit den ihm bekannten Werkzeugen Objektteile an den orthographischen Ansichten auszurichten [Abbildung 1].

Abbildung 1 *(von links nach rechts)*: Bilder der Eingangsbildfolge, erzeugte orthographische Ansichten, mit Hilfe der orthographischen Ansichten interaktiv modelliertes 3D-Oberflächenmodell

Abbildung 2: Eingangsbildfolgen mit überlagertem 3D-Modell einer bewegten Person *(in gelb)*, sowie die 3D-Rekonstruktion der gesamten Szene inklusive Halfpipe und bewegten Kameras *(rechts)*

Bewegungsschätzung einer Person mit nicht synchronisierten bewegten Kameras

In diesem Projekt wurde ein neuer Ansatz zur Bewegungsschätzung von Personen *(Markerless Motion Capture)* entwickelt, wobei die bewegten Personen mit mehreren bewegten und nicht-synchronisierten Kameras aufgenommen wurden, anstelle der üblichen feststehenden und synchronisieren Kameras. Dieser Ansatz erlaubt es, die Bewegungsschätzung von Personen mit günstigen in der Hand gehaltenen Videokameras durchzuführen. Zur Vorbereitung einer Sequenz wird zunächst der statische 3D-Hintergrund der Szene und die Position der einzelnen Kamera mit einem so genannten Structure-and-Motion (SaM) Verfahren ermittelt. Dann werden die Kameras mit Hilfe der rekonstruierten Geometrie des statischen Hintergrunds zueinander registriert. Die Synchronisation der Kameras erfolgt über die Audiospur, die von den Kameras parallel aufgezeichnet wird. Schließlich wird ein Verfahren zur Bewegungsschätzung angewendet, um die Position und die Gelenkpositionen der Person zu bestimmen [Abbildung 2]. Verfolgte Merkmalspunkte in der Bildfolge sowie die rekonstruierte Geometrie des statischen Hintergrunds können zur weiteren Stabilisierung der Bewegungsschätzung herangezogen werden.

Detektion von Gesichtern mit Hilfe von Stereokameras

In diesem Projekt wurde ein Verfahren zur Detektion von Gesichtern in Bildfolgen untersucht. Es verbessert den Stand der Technik für 2D-Objektdetektion durch die Auswertung einer zusätzlichen Disparitätskarte, die für eine Bildregion mit Hilfe eines kalibrierten Stereo-Kameraaufbaus geschätzt wird. Dabei werden die Gesichter zunächst in den 2D-Bildern detektiert. In einem zweiten Schritt werden fälschlicherweise detektierte Gesichter durch die Analyse der Disparitätskarte eliminiert.

Diese neuartige Kombination von Algorithmen benötigt eine sehr geringere Rechenzeit und reduziert die Anzahl der falsch detektierten Gesichter im Vergleich zu klassischen 2D-Gesichtsdetektionsverfahren. :::

Abbildung 3:
a) linkes Bild der Eingangsbildfolge
b) rechtes Bild mit detektierten Gesichtern
c) Disparitätskarte
d) mit Hilfe der Disparitätskarte wird das linke Gesicht als falsch erkannt

KONTAKT

Thorsten Thormählen
ABT. 4 Computergrafik
Telefon +49 681 9325-417
Email thormae@mpi-inf.mpg.de

Korrespondenzen und Symmetrien in 3D-Objekten

Abbildung 1: Ein Scan eines Gebäudes
(Neues Rathaus in Hannover, Daten zur Verfügung gestellt vom IKG, Universität Hannover)

Abbildung 2: Die automatische Zerlegung des Gebäudes in Bausteine. Gleiche Typen von Bausteinen sind in gleichen Farben dargestellt, gegebenenfalls mit gestrichelten Trennlinien.

Einleitung

Datenbanken geometrischer Objekte werden in der Wissenschaft wie auch im täglichen Leben immer wichtiger. Heute werden in der medizinischen Praxis routinemäßig 3D-Scans von Patienten gemacht (mit Röntgenstrahlen, Magnetresonanz oder Positronenemission), die wichtige Einblicke in Abläufe in unserem Körper geben, vom gebrochenen Knochen bis zur Gehirnaktivität. In der Biologie gibt es mittlerweile Verfahren (z.B. Elektronentomographie), die dreidimensionale Abbildungen von Proteinen bis hin zu ganzen Zellen in molekularer Auflösung ermöglichen. Auch in unserem täglichen Leben haben wir uns daran gewöhnt mit Anwendungen wie Google Earth oder Microsoft Virtual Earth unsere Lebensumgebung zu erkunden. Hier werden zurzeit von verschiedenen Seiten Großprojekte im industriellen Maßstab durchgeführt, um hochauflösende 3D-Scans großer Gebiete zu erfassen. Dies bedeutet: 3D-Scans aller wichtigen Ballungszentren (und mehr) dieses Planeten mit hohen Auflösungen werden damit voraussichtlich in allernächster Zeit verfügbar werden.
Angesichts dieser Entwicklungen wird es immer wichtiger, diese Daten, die in großer Menge anfallen, automatisch zu organisieren. Gleichzeitig ist es auch aus wissenschaftlicher Sicht immer interessanter, in solchen Daten automatisch nach auffälligen Mustern und Strukturen zu suchen. Dies hat breite Anwendungen von der medizinischen Diagnose aus Tomographiescans (gelernt aus vielen Beispielen) bis hin zur automatischen Modellierung von Gebäuden nach gelernten Mustern im Bereich der reinen Computergrafik.

Korrespondenzen und Symmetrien

Diese Entwicklungen motivieren dazu zu untersuchen, wie Muster in geometrischen Daten automatisch gefunden werden können. Das Fernziel wäre allgemeine *„Strukturen"* automatisch zu *„verstehen"*. Ein solches Problem kann natürlich zurzeit nicht auf dem kognitiven Niveau menschlicher Wahrnehmung gelöst werden; die aktuelle Forschung konzentriert sich daher hier zunächst einmal auf elementare und grundlegende Probleme. Ein wichtiger Schritt dabei ist die Lösung des Korrespondenzproblems: Hier ist die Frage, wann zwei geometrische Objekte, oder Teile davon, im Wesentlichen identisch sind. Je nachdem, was man unter *„im Wesentlichen"* versteht, ergeben sich hier verschiedene Problemstellungen: Die einfachste Frage ist festzustellen, ob zwei Objekt(teile) bis auf eine starre Bewegung (Rotation, Verschiebung) identisch sind. Die Frage wird komplexer, wenn zugelassen wird, dass sich die Objekte deformieren können (zum Beispiel ein 3D-Scan einer Person in verschiedenen Posen). Noch komplizierter ist die Definition allgemeiner Korrespondenzen: Wie erkennt man, dass zwei Autos ähnlich sind, obwohl die geometrische Erscheinung sehr verschieden ist. Ein spezielle Korrespondenzproblem ist die Symmetrieerkennung: Hier wird untersucht, ob ein Objekt vollständig oder teilweise zu sich selbst ähnlich ist. Man kann dies so verstehen, dass man *„Bausteine"* berechnen möchte, aus denen das Objekt aufgebaut ist, ohne diese a priori zu kennen. Dies liefert Aufschluss über Struktureigenschaften eines geometrischen Objektes rein aus gemessenen Daten, ohne Vorabinformationen über die gefundenen Bauteile zu benötigen.

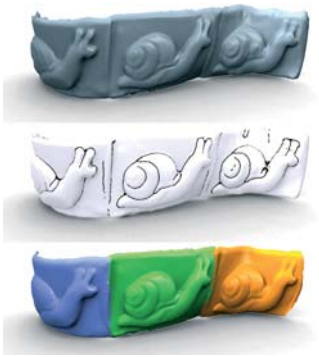


Abbildung 3: Ein komplexeres Beispiel, bei dem die Bausteine stark deformiert sind. Um sie dennoch erkennen zu können, werden die Graphen von Linienmuster auf der Oberfläche fehlertolerant verglichen.

Einige Beispiele

Die Abbildungen zeigen Beispiele aus aktuellen Arbeiten. Das erste Bild zeigt die Zerlegung eines Hauses [Abbildung 1] in starre Bausteine [Abbildung 2]. Eine allgemeinere Zerlegung ist in Abbildung 3 zu sehen: In der Figur, die von einem Plastilinmodell gescannt wurde, sind mehrere ähnliche Figuren zu sehen, die zueinander stark deformiert wurden. Dennoch konnte automatisch erkannt werden, dass es sich um ein und dasselbe Muster handelt. Neben der Erkennung von Bausteinen in 3D-Scans ergeben sich weitere Anwendungen, zum Beispiel in der Verarbeitung von 3D-Filmen, die mit Echtzeitscannern erfasst wurden, oder der Mustererkennung in 2D-Bildern [Abbildung 4]. :::

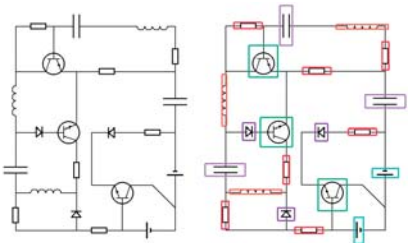


Abbildung 4: Das Symmetriedetektionsverfahren eignet sich auch für andere Typen von Daten, wie hier ein Bitmap-Bild eines Schaltplans, in dem Bauteile automatisch und ohne Vorwissen erkannt werden.

KONTAKT

Michael Wand
ABT. 4 Computergrafik
Telefon +49 681 9325-408
Email mwand@mpi-inf.mpg.de

Dreidimensionale Animation und Rekonstruktion von Gesichtern

Synthetisch generierte Szenen fanden in den letzten Jahren häufig Einsatz in Film- und Fernsehproduktionen. Zunächst beschränkten sich die am Computer generierten Szenen vorwiegend auf aktionsreiche Szenen. Im Laufe der Jahre haben sich die Techniken, mit denen dreidimensionale Modelle animiert und abgebildet werden können stetig verbessert. Das führte dazu, dass heute nahezu jede Szene eines Films am Computer nachbearbeitet wird. Dennoch werden dreidimensionale Objekte bis heute häufig noch von Hand erzeugt und in mühsamen Einzelschritten von erfahrenen 3D-Spezialisten am Rechner animiert, beleuchtet und mit real gedrehten Szenen verknüpft.

Ein Ziel der Arbeitsgruppe ist es, diesen Arbeitsprozess zu automatisieren und die Ergebnisse zu verbessern. Schnelle, exakte und automatische Methoden zu finden, die dreidimensionale Modelle vergleichbar detailgetreu, realistisch animieren und abbilden können. Im Speziellen beschäftigt sich die Gruppe dabei mit der Animation von Gesichtern mit Hilfe so genannter lernbasierter Methoden.

Bei lernbasierten Ansätzen wird in der Regel zunächst eine große Beispiel-Datenmenge gesammelt, aus der mit Hilfe mathematischer Methoden dann objektspezifische Eigenschaften extrahiert beziehungsweise erlernt werden können. Um dreidimensionale Gesichtsdaten zu sammeln, benutzten die Mitglieder der Gruppe schnelle 3D-Scanner die mit Hilfe eines Laserstrahls oder eines Beleuchtungsmusters die Oberfläche eines Gesichts in 3D vermessen. Während die Laser-Scanner verhältnismäßig langsam arbeiten, dafür aber sehr exakt, lassen sich mit beleuchtungsbasierten 3D-Scannern schnelle Mimikbewegungen erfassen, auf Kosten der Detailtreue. Die Kombination beider Ansätze bietet ein breites Spektrum an Anwendungsmöglichkeiten.

a b c d

Mit Hilfe eines Beleuchtungsmodells konnten starke Beleuchtungsunterschiede in den Eingabebildern (a,c) so reduziert werden, dass die Texturen der zugehörigen 3D-Rekonstruktionen (b,d) weitestgehend von Überbelichtung und Schlagschatten befreit sind.

Lippensynchrone Animation von Gesichtern in 3D

Zum Beispiel wurde eine neue Methode entwickelt, die in Gesichtern künstliche Mundbewegungen erzeugen kann. Aus einer Reihe von dreidimensionalen Bewegtbild-Aufnahmen von Gesichtern konnten Mitglieder der Arbeitsgruppe mit Hilfe statistischer Methoden lernen, wie sich die Lippen eines Gesichts beim Sprechen bewegen. Das so erlernte Wissen kann auf neue, unbekannte Gesichter übertragen werden. Dazu benötigt man als Eingabe nur eine einfache Audio-Datei. Was bislang nur mühsam und zeitaufwendig von Hand animiert werden konnte, ist mit Hilfe der neuen Technik in Sekundenschnelle erledigt. Animations-Spezialisten spart diese Technik Zeit, und ermöglicht ihnen, verschiedene Ansätze auszuprobieren.

Automatische Vervollständigung von lückenhaften 3D-Scans

Ein weiteres Ziel der Arbeitsgruppe ist es, die Qualität der aufgenommenen Daten stetig zu verbessern. Unter anderem spielen 3D-Scans heute eine große Rolle bei der Gesichtserkennung. Dazu werden in 3D erfasste Gesichter mit einer Vielzahl von Fotos oder Scans verglichen, um eine Übereinstimmung zu finden. Trotz der exakten Messmethode der Laser-Scanner passiert es jedoch häufig, dass glänzende Oberflächen oder Verdeckungen (z.B. durch Haare) bei der Aufnahme zu lückenhaften Ergebnissen führen. Sind die so aufgenommenen 3D-Daten unvollständig oder schlecht ausgeleuchtet, wird ein Vergleich mit Fotos zunehmend schwieriger.

Um lückenhafte 3D-Laser-Scans zu vervollständigen und von starken Beleuchtungseffekten zu befreien stellten Mitglieder der Arbeitsgruppe ein neues Verfahren vor. Dabei nutzten sie die Erkenntnisse, die sie aus etwa 500 3D-Scans von Gesichtern zogen. Mit Hilfe der großen Datenbasis war es möglich die fehlenden Bereiche im neu erfassten Scan zu rekonstruieren und detailgetreu nachzubilden. Auch starke Beleuchtungseffekte konnten mit Hilfe eines Beleuchtungsmodells rückgängig gemacht werden [Abbildung]. So konnten die neu erfassten 3D-Scans zuverlässig und robust mit Fotos oder anderen Scans verglichen werden. Exemplarisch zeigten die Mitglieder der Gruppe, dass Gesichtserkennungsprogramme ein deutlich höheres Leistungsvermögen haben können, wenn sie anstelle von Fotos dreidimensionale Daten verwenden.

Da 3D-Scanner im Laufe der letzten Jahre immer günstiger wurden, ist es denkbar, dass vergleichbare Techniken zukünftig vermehrt im Alltag Einsatz finden werden. :::

KONTAKT

Kristina Scherbaum
ABT. 4 Computergrafik
Telefon +49 681 9325-454
Email scherbaum@mpi-inf.mpg.de

INFORMATIONSSUCHE & DIGITALES WISSEN

Digitale Information hat unsere Gesellschaft und Wirtschaft, das Arbeiten in den Wissenschaften und das Alltagsleben fundamental verändert. Moderne Suchmaschinen liefern zu praktisch jeder Frage nützliche Information, und das Internet hat das Potential, der Welt umfassendste Sammlung maschinell verarbeitbaren Wissens zu sein. Doch Wissensstrukturen im Internet sind amorph, und Suchmaschinen haben selten präzise Antworten auf Expertenfragen, für die man Lexika und Fachliteratur zu Rate ziehen muss. Eine große Herausforderung und Chance ist der Schritt vom Rohstoff Information zum computergestützten, intelligenten Umgang mit digitalem Wissen.

Parallel zum Anvisieren dieses Quantensprungs beobachten wir eine Komplexitätsexplosion beim Rohstoff digitaler Informationen entlang verschiedener Dimensionen: Quantität, strukturelle Vielfalt, Multimodalität, digitale Historie und Verteilung.

■ Zusätzlich zu den mehr als 20 Milliarden Webseiten zählen heute Online-Nachrichtenströme, Blogs, Tweets und soziale Netze mit mehreren hundert Millionen von Benutzern, Web2.0-Communities über Photos, Musik, Bücher, wissenschaftliche Spezialthemen und nicht zuletzt die Enzyklopädie Wikipedia zu den potentiell wichtigen Informationsquellen. Das Gesamtvolumen dieser Daten liegt in der Größenordnung von Exabytes: 10 hoch 18 Bytes – mehr als eine Millionen Terabyte-Platten.

■ Dabei kommen zunehmend ausdrucksstärkere Datenrepräsentationen zum Einsatz: XML-Dokumente, RSS-Feeds, semantisch verknüpfte RDF-Graphen und vieles mehr. Die reichere Struktur und Heterogenität der Daten erhöht wiederum die Komplexität zur Beherrschung dieser digitalen Vielfalt.

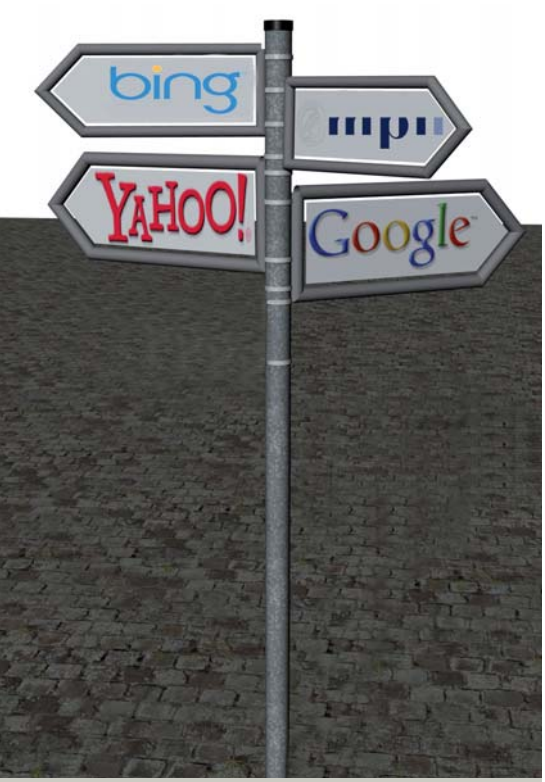
■ Zusätzlich zu textorientierten und strukturierten Daten erleben wir eine Explosion multimodaler Information: Milliarden von Menschen werden zu Datenproduzenten im Web, indem sie ihre Bilder, Videos und Tonaufzeichnungen mit dem Rest der Welt teilen. Dies geht häufig einher mit zwischenmenschlichen Kontakten, die über das Internet entstehen und in großen Online-Netzen organisiert sind.

■ Die Historie digitaler Information – beispielsweise frühere Versionen unserer Instituts-Webseite, die zum Teil vom Internet Archive konserviert werden – ist eine potentielle Goldmine für tiefergehende Analysen entlang der Zeitdimension. Davon können Soziologen und Politologen profitieren, aber auch Medien- und Marktanalysten sowie Experten für geistiges Eigentum.

■ Die Quantität und Vielfalt der im Internet verfügbaren Information ist so hoch geworden, dass Suchmaschinen längst nicht mehr alle relevanten Verweise in einem zentralen Index vorrätig halten können. Daher muss globale Informationssuche langfristig mit verteilten Algorithmen angegangen werden, indem beispielsweise viele lokale Suchmaschinen für spezifische Aufgaben dynamisch föderiert werden.Hier spielen dann nicht nur die lokale Rechen- und Suchgeschwindigkeit eine wichtige Rolle, sondern auch die Kommunikationseffizienz im Netz der Netze, dem Internet, und den darin eingebetteten Peer-to-Peer-Netzen.

Am Max-Planck-Institut für Informatik wird dieses globale Thema unter verschiedenen Blickwinkeln untersucht. Dazu gehören die effiziente Suche auf semistrukturierten XML-Dokumenten, die vor allem in digitalen Bibliotheken und bei e-Science-Daten eine wichtige Rolle spielen, und der skalierbare Umgang mit graphstrukturierten RDF-Daten, die im Semantic-Web-Kontext entstehen, aber auch als Datenrepräsentation in der Computational-Biology an Bedeutung gewinnen. Bei anderen Projekten steht die benutzerorientierte Sicht auf Web2.0-Communities und multimodale Daten im Vordergrund.Die große Vision vom Quantensprung zur Wissenssuche schließlich wird in Arbeiten über automatische Wissensextraktion aus Web-Quellen wie zum Beispiel Wikipedia verfolgt. Das Max-Planck-Institut für Informatik hat hier eine weltweite Vorreiterrolle. :::

BEITRÄGE



Intelligente und trotzdem schnelle Suche

Suchmaschinen sind längst ein unentbehrliches Werkzeug in den verschiedensten Bereichen geworden. Wissenschaftler suchen nach Literatur zu einem bestimmten Thema, eine Studentin sucht im Katalog der Universitätsbibliothek, eine Schülerin nach Material für ein Referat, Krankenhäuser suchen in Unmengen von Patientendaten, Juristen in Fällen, fast jeder in seiner Mail ... und die Liste ließe sich fast endlos fortsetzen.

Die meisten Suchmaschinen sind heutzutage sehr schnell: wir haben uns längst an Antwortenzeiten im Untersekundenbereich gewöhnt, egal ob Millionen oder Milliarden von Dokumenten durchsucht werden. Die meisten Suchmaschinen sind allerdings nicht sehr schlau: sie machen einfach Stichwortsuche, das heißt, sie finden diejenigen Dokumente, die die Stichworte enthalten die man eingegeben hat.

Das funktioniert oft sehr gut, zum Beispiel gelangt man so mit der Suchanfrage *„uni saarland“* sofort auf die Hauptseite der Universität, einfach weil sie diese Worte prominent in Titel und Webadresse enthält. Überhaupt nicht funktioniert dagegen Stichwortsuche bei einer Anfrage wie *„wissenschaftler saarbrücken“*. Man gebe das einfach mal in eine Suchmaschine wie Google ein. Erwarten würde man eine Seite mit den Namen von Wissenschaftlern die in Saarbrücken forschen, oder wenigstens Seiten von solchen Wissenschaftlern. Und jeder Wissenschaftler und jede Wissenschaftlerin hat ja ihre eigene Webseite. Da steht aber nicht das Wort Wissenschaftler, und deswegen wird sie nicht gefunden.

Ein Problem dabei: Maschinelles Lernen

Das genannte Problem hat zwei Komponenten. Zum einen müssen wir der Suchmaschine beibringen, was ein Wissenschaftler ist bzw. welche Zeichenfolgen auf Webseiten einen Wissenschaftler benennen. Das ist ein Problem des *maschinellen Lernens*. Ein einfacher Algorithmus lässt sich an folgenden Beispiel illustrieren. Es gibt Albert Einstein, den berühmten Physiker, und Alfred Einstein, den auch bekannten aber nicht ganz so berühmten Musikologen. Nun ist auf einer Webseite nur von *„Einstein“* die Rede. Wie finden die Maschinen heraus, welcher Einstein gemeint ist? Nun steht ja auf vielen Webseiten der volle Name und wir wissen welcher Einstein gemeint ist. Schauen wir uns die Worte in der unmittelbaren Umgebung des Namens an, finden wir bei Albert Einstein unverhältnismäßig oft Worte wie *„Physik“* und *„Relativitätstheorie“*, bei Alfred Einstein dagegen eher *„Musik“* und *„Brahms“*. Finden wir nun in der unmittelbaren Umgebung eines Vorkommens von *„Einstein“* (ohne Vornamen) das Wort *„Relativitätstheorie“* können wir mit hoher Wahrscheinlichkeit davon ausgehen, dass Albert Einstein gemeint war.

Das zweite Problem: Effiziente Suchstrukturen

Nehmen wir nun an, die Maschine hat all dieses Wissen erworben. Wie schafft sie es nun, im Bruchteil einer Sekunde aus Milliarden von Seiten die herauszufiltern, wo von einem Saarbrücker Wissenschaftler die Rede ist?

Für reine Stichwortsuche ist dieses Problem recht einfach. Wir berechnen einfach für jedes mögliche Wort die Liste aller Dokumente vor, die dieses Wort enthalten. So wissen wir, wenn eine Suchanfrage kommt, auf Anhieb, dass das Wort *„uni“* in den Seiten mit den Nummern 17, 34, 118, usw. vorkommt und das Wort *„saarland“* in den Dokumenten mit den Nummern 23, 34, 132, usw. Die Nummern zu finden, die in beiden Listen vorkommen ist nun für eine Maschine ein leichtes, und so erhalten wir selbst bei riesigen Dokumentensammlungen die Ergebnisse extrem schnell.

Soll komplexere Information gespeichert werden, wie z.B. für jedes Vorkommen von Albert Einstein die Nennung aller seiner Eigenschaften, versagt dieser einfache Ansatz. Traditionell kommen an dieser Stelle Datenbanken zum Einsatz. Datenbanken sind sehr mächtige Programme, die auf fast beliebig großen Datenmengen auch komplexeste Operationen ausführen können. Der Preis dafür ist allerdings eine gewisse Schwerfälligkeit: Würde man eine Suchmaschine als Datenbankanwendung programmieren, wäre sie etwa 1000-mal langsamer.

CompleteSearch: das beste aus beiden Welten

Die am Max-Planck-Institut für Informatik entwickelte *CompleteSearch Technologie* vereint nun in gewisser Weise das beste aus beiden Welten. CompleteSearch ist genauso schnell wie eine herkömmliche Suchmaschine, unterstützt aber weitaus komplexere Operationen, wie zum Beispiel die im zweiten Absatz beschrieben semantische Suche *„wissenschafter saarbrücken“*.

Möglich wurde dies durch zwei Entdeckungen. Zum einen ist dies die Formalisierung des Problems der so genannten *„kontext-sensitiven Präfixsuche“*, das einerseits einfach genug ist, dass man es effizient lösen kann, und zwar genauso effizient wie das Standard Suchmaschinen Problem. Zum anderen haben wir eine Lösung für diese Art Präfixuche entwickelt, die sowohl in der Theorie beweisbar effizient als auch in der Praxis rasend schnell ist.

Auf der Webseite http://search.mpi-inf.mpg.de stehen eine ganze Reihe von Demonstrationen der CompleteSearch Suchtechnologie zum Ausprobieren zur Verfügung. :::

KONTAKT

Hannah Bast
ABT. 1 Algorithmen und Komplexität
Telefon +49 681 9325-120
Email bast@mpi-inf.mpg.de

Zufällige Telefonketten – effiziente Kommunikation in Datennetzen

Wie kann man schnell und effizient eine eilige Nachricht an eine größere Gruppe von Menschen schicken? Ein klassisches Verfahren ist die Telefonkette. Im einfachsten Fall basiert sie auf einer geordneten Liste der Gruppenmitglieder. Die zu verbreitende Neuigkeit wird zunächst dem ersten auf dieser Liste mitgeteilt. Der ruft nun den zweiten an, der den dritten, und so weiter.

In der Informatik treten ähnliche Aufgabenstellungen auf, wenn Informationen in Rechnernetzen verbreitet werden müssen. Ein Unternehmen mit Filialen an mehreren Standorten könnte seine Kundendaten in jeder Filiale lokal speichern. Dies hat den Vorteil, dass in jeder Filiale schnell und unkompliziert auf diese Daten zugegriffen werden kann, und dass die Daten auch bei kurzzeitiger Nichterreichbarkeit der zentralen Datenbank zur Verfügung stehen. Bei Verwendung solcher replizierter Datenbanken müssen natürlich Änderungen, die sich im Datenbestand einer Filiale ergeben, zügig an alle anderen kommuniziert werden.

In zwei zentralen Aspekten unterscheidet sich dieses Telefonkettenproblem der Informatik von dem der Alltagswelt. Erstens können solche Datenbanksysteme sehr groß sein. Bei einem Netz von einigen tausend Knoten würde die Informationsausbreitung viel zu lange dauern, wenn die Knoten in einer festen Reihenfolge nacheinander die Nachricht weitergeben.

Ein zweiter Punkt ist die Robustheit des Verfahrens. Unter Robustheit versteht man, dass ein Verfahren auch dann noch gut funktionieren soll, wenn einzelne Teilschritte nicht ganz so abgelaufen sind, wie man es erhofft hat. Bei einer Telefonkette ist dies hauptsächlich das Problem, dass ein Teilnehmer nicht erreichbar ist, oder dass er zwar erreichbar ist, aber dann aus irgendwelchen Gründen die Nachricht nicht weitergibt. Die klassische sequentielle Telefonkette ist augenscheinlich nicht sehr robust. Sowie ein Teilnehmer die Nachricht nicht weitergibt, bleiben alle nachfolgenden uninformiert.

Aus diesen Gründen sind Verfahren, die sich an der klassischen Telefonkette orientieren, für die Anwendung in der Informatik ungeeignet. Dennoch gibt es für die in der Informatik auftretenden Telefonkettenprobleme eine überraschend einfache Lösung. Bei der *zufälligen Telefonkette* ruft jeder, der die Nachricht kennt, zufällig gewählte andere Teilnehmer an. Zur Vereinfachung der Darstellung sei angenommen, dass alle Anrufe gleich lange benötigen. Dies führt dazu, dass der Informationsaustausch in Runden abläuft. In jeder Runde ruft jeder informierte Teilnehmer bei einem zufällig gewählten anderen Teilnehmer an. Dieser ist dann spätestens ab diesem Zeitpunkt ebenfalls informiert.

Die zufällige Telefonkette ist erstaunlich effizient und robust. Nehmen wir als Beispiel ein Datennetz mit 1024 Knoten, die alle miteinander kommunizieren können. Dann benötigt die zufällige Telefonkette im Schnitt nur 18,09 Runden, bis alle Knoten informiert wurden. Ähnlich gut sieht es mit der Robustheit aus. Selbst wenn wir annehmen, dass jeder zehnte Teilnehmer nie erreichbar ist, genügen im Schnitt nur 19,49 Runden, um die übrigen zu informieren. Dabei spielt es keine Rolle, welche 10% der Teilnehmer ausfallen.

Diese positiven Eigenschaften motivieren eine tiefergehende Untersuchung von derartigen randomisierten Protokollen. Spannend ist insbesondere die Frage, was die richtige Dosis von Zufall ist. Aktuelle Ergebnisse suggerieren, dass eine Kombination von zufälligen Elementen mit dem klassischen Listenverfahren die besten Resultate liefert. :::

KONTAKT

Benjamin Doerr
ABT. 1 Algorithmen und Komplexität
Telefon +49 681 9325-104
Email doerr@mpi-inf.mpg.de

Informationssuche in Web-Archiven

Das World-Wide-Web (kurz: Web) wächst ständig, und täglich kommen neue Inhalte hinzu. Ein Teil dieser Inhalte wird im Web erstmals und ausschließlich veröffentlicht und spiegelt aktuelles Geschehen wieder. In den letzten Jahren ist das Bewusstsein gewachsen, dass im Web veröffentlichte Inhalte wertvoll sind und langfristig bewahrt werden müssen. Nationalbibliotheken und Organisationen wie das Internet Archive (http://www.archive.org) haben diese Aufgabe übernommen. Andere Inhalte wurden ursprünglich vor langer Zeit veröffentlicht und sind nun erstmals, dank verbesserter Digitalisierungsverfahren, im Web verfügbar. Ein Beispiel hierfür sind Zeitungsarchive. Das im Web zugängliche Archiv der britischen Zeitung *The Times* etwa reicht bis ins Jahr 1785 zurück.

Damit Web-Archive für Benutzer leicht zugänglich und für Analysen nützlich sind, bedarf es ausgeklügelter Suchverfahren. Diese sind Gegenstand unserer gegenwärtigen Forschung. Im Folgenden beschreiben wir drei Teilaspekte unserer Arbeiten.

Zeitreisen in Web-Archiven

Bestehende Suchverfahren ignorieren die in Web-Archiven vorhandene Zeitdimension. So ist es beispielsweise nicht möglich, eine Suche nur auf jene Dokumente zu beschränken, die in einem gewissen Zeitraum oder zu einem bestimmten Zeitpunkt existiert haben. Unter solch einer Zeitreise-Anfrage verstehen wir eine aus Schlüsselwörtern bestehende Anfrage wie *„Prognosen zur Bundestagswahl“*, die um einen zeitlichen Kontext, beispielsweise September 2009, erweitert ist. Ergebnis dieser Anfrage sind relevante Dokumente, die im genannten Zeitraum tatsächlich existiert haben. Unsere Verfahren basieren auf dem invertierten Index, welcher für jedes Wort eine Liste mit Informationen zu den Vorkommen des Wortes in einzelnen Dokumenten enthält. Wir erweitern die pro Vorkommen gespeicherte Information um ein Gültigkeits-Zeitintervall. Zudem nutzen wir aus, dass sich verschiedene Versionen eines Dokumentes typischerweise nur wenig unterscheiden. Dies erlaubt uns, die Größe des Index dramatisch zu reduzieren. Eine Beschleunigung der Anfragebearbeitung lässt sich durch redundante Datenhaltung erzielen, indem man für jedes Wort mehrere Listen mit den Wortvorkommen in bestimmten Zeiträumen unterhält. Daraus ergibt sich ein Zielkonflikt zwischen Zeiteffizienz und Platzbedarf. Unser Ansatz beinhaltet verschiedene Optimierungsverfahren, um in diesem Zielkonflikt unter bestimmten Vorgaben (beispielsweise einer Beschränkung des Platzbedarfs) zu vermitteln.

Umgang mit veränderter Terminologie

Sprachgebrauch und Terminologie wandeln sich ständig. Ein prominentes Beispiel hierzu ist die Stadt Sankt Petersburg, welche bis 1991 als Leningrad bekannt war. Die ständige Veränderung von Terminologie stellt eine Herausforderung für die Informationssuche auf Web-Archiven dar. Manche der archivierten Dokumente wurden vor langer Zeit veröffentlicht (z.B. im späten 18. Jahrhundert) und bedienen sich der Terminologie dieser Zeit. Benutzer jedoch formulieren ihre Suchanfragen unter Verwendung heute gängiger Terminologie. Bildlich gesprochen liegt damit eine sich ständig weitende Kluft zwischen den archivierten Dokumenten und den Suchanfragen heutiger Benutzer, wie das folgende Beispiel illustriert. Ein Kunstliebhaber, der an Museen in Sankt Petersburg interessiert ist, stellt die Suchanfrage *„museum sankt petersburg“*. Bei Verwendung existierender Suchverfahren werden Dokumente, die vor 1991 veröffentlich wurden und detailliert über Museen in Leningrad berichten, nicht gefunden. Diese älteren aber dennoch relevanten Dokumente bleiben unserem Kunstliebhaber damit verborgen.

Unsere Verfahren formulieren die Suchanfrage des Benutzers automatisch so um, dass auch ältere aber relevante Dokumente gefunden werden. Zuerst identifizieren wir Wörter, die in der Vergangenheit eine sehr ähnliche Bedeutung wie die Wörter in der Suchanfrage des Benutzers hatten. Daraufhin versuchen wir, die so gefundenen Wörter derart zu kombinieren, dass eine sinnvolle Suchanfrage entsteht, die das Informationsbedürfnis des Benutzers widerspiegelt. Beides geschieht basierend auf zeitbezogenen Textstatistiken, die wir für unser Web-Archiv berechnen.

Informationsbedürfnisse mit Zeitbezug

Informationsbedürfnisse von Benutzern haben häufig einen Zeitbezug. Dieser kann unmittelbar aus der Suchanfrage ersichtlich sein, beispielsweise wenn diese explizit ein Jahr oder Jahrhundert erwähnt. Web-Archive sind ideale Dokumentensammlungen, um solche Informationsbedürfnisse zu bedienen. Bestehende Suchverfahren scheitern jedoch oft an diesen zeitbezogenen Informationsbedürfnissen. Der Grund hierfür ist, dass ihnen die Bedeutung in Dokumenten enthaltener Zeitbezüge verborgen bleibt. Für die Suchanfrage *„deutsche maler 15. jahrhundert“* wird ein Dokument mit Details zum Leben von Albrecht Dürer, welches viele Jahresangaben wie sein Geburtsjahr 1471 enthält, nicht zwingend als relevant eingestuft. Der Grund hierfür ist, dass bestehenden Suchverfahren nicht wissen oder schließen können, dass das Jahr 1471 im 15. Jahrhundert liegt.

Um solche Beziehungen zwischen Zeitbezügen zu erkennen, repräsentieren unsere Verfahren Zeitbezüge formal als Zeitintervalle. Dieses Wissen kann dann verwendet werden, um bessere Suchergebnisse für Informationsbedürfnisse mit Zeitbezug zu erreichen. Hierzu integrieren unsere Verfahren das Wissen über Zeitbezüge in so genannte *Statistical Language Models*. :::

KONTAKT

Klaus Berberich
ABT. 5 Datenbanken und Informationssysteme
Telefon +49 681 9325-528
Email kberberi@mpi-inf.mpg.de

Informationssuche in sozialen Netzen

Das Aufkommen des Web 2.0 hat eine Revolution im Umgang mit dem Web ausgelöst.

Statt Information nur zu konsumieren, kann nun jedermann einfach selbst Inhalte generieren und veröffentlichen. Onlinedienste wie del.icio.us, Flickr, LibraryThing, YouTube, MySpace, Facebook oder die vor allem in Deutschland verbreiteten Dienste StudiVZ und werkennt-wen bieten aber nicht nur Speicherplatz für Bilder, Videos, Bookmarks und ähnliche Daten, sondern erlauben auch Interaktion: Benutzer können Daten mit anderen Benutzern teilen, die Daten anderer Benutzer erforschen, kommentieren, bewerten und sie mit Schlagwörtern, so genannten Tags, versehen. Das manuelle Annotieren von Inhalten – der eigenen, aber auch der anderer Benutzer – ist dabei eine der wichtigsten Funktionen. Die so vergebenen Tags sind oft sehr gute inhaltliche Beschreibungen, weil sie frei gewählt werden können, ohne sich an ein vorgegebenes Schema halten zu müssen. Verschiedene Benutzer belegen häufig das gleiche Bild oder das gleiche Video mit unterschiedlichen Tags, die ihre unterschiedlichen Interessen wiedergeben. Die meisten Dienste bieten komfortable und intuitive Schnittstellen, um Inhalte basierend auf ihren Annotationen zu finden, zum Beispiel über *„Tagwolken"*.

Zusätzlich erlauben es die Dienste explizite Listen von Freunden zu unterhalten und bieten oft einen damit zusammenhängenden Mehrwert, zum Beispiel Freunde automatisch über neue Inhalte zu informieren. Auf diese Weise bildet sich ein *soziales Netz*, ein dichtes Beziehungsgeflecht der Benutzer, in dem die Anzahl der Freunde eines Benutzers oft als Indikator für seinen Ruf gesehen wird. Während die Liste der Freunde initial mit Freunden und Bekannten aus dem „echten" Leben gefüllt wird, die den gleichen Dienst nutzen, wächst sie im Lauf der Zeit um vorher unbekannte Benutzer, die ähnliche Interessen verfolgen.

Die dichte Vernetzung der Benutzer und die vielfältigen Annotationen zusammen erlauben es die „Weisheit der Massen" auszunutzen, um wertvolle Inhalte zu finden, die von Freunden empfohlen werden - entweder von direkten Freunden oder transitiv von Freunden der eigenen Freunde, und entweder explizit (z.B. durch Bewertungen und Kommentare) oder implizit (z.B. durch die intensive Vergabe von Annotationen). Bei der Suche in solchen Systemen sollten daher die Beziehungen zwischen Benutzern berücksichtigt werden, da man in der Regel seinen engen Freunden stärker vertraut als flüchtigen Bekannten, also solchen Benutzern, die im Beziehungsnetz weit entfernt sind. In den vorhandenen Systemen sucht man solche Funktionen bisher aber weitgehend vergebens.

Die von uns entwickelte Suchmaschine *SENSE („Socially ENhanced Search and Exploration")* schließt diese Lücke durch eine Suchfunktion über Inhalten in sozialen Netzen, die Annotationen von Benutzern je nach der Stärke ihrer Beziehung zu dem Benutzer gewichtet, der die Suchanfrage gestellt hat. Neben dem Abstand der Benutzer im Beziehungsnetz kann außerdem die inhaltliche Überlappung der Annotationen, die sie verwendet haben, die Gewichtung beeinflussen. Im Vergleich zu den in heutigen Systemen vorhandenen, rein häufigkeitsbasierten Suchen lässt sich durch eine solche personalisierte Suche eine signifikante Steigerung der Ergebnisqualität erzielen. SENSE verwendet dabei hocheffiziente und skalierbare Suchalgorithmen, um mit dem schnellen Wachstum dieser Dienste und der sehr hohen Rate, mit der neue Inhalte und Annotationen generiert werden, zurechtzukommen. :::

KONTAKT

Tom Crecelius
ABT. 5 Datenbanken und Informationssysteme
Telefon +49 681 9325-506
Email tcrecel@mpi-inf.mpg.de

Ralf Schenkel
ABT. 5 Datenbanken und Informationssysteme
Telefon +49 681 9325-504
Email schenkel@mpi-inf.mpg.de

YAGO – eine digitale Wissensammlung

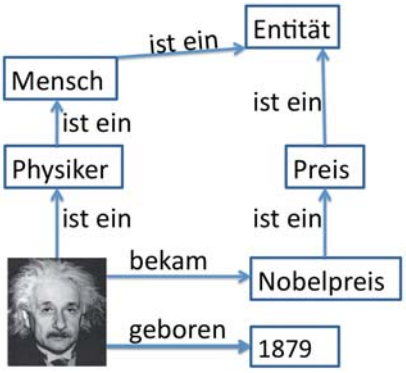
In den letzten Jahren hat sich das Internet zu einer bedeutenden Informationsquelle entwickelt. Zugfahrpläne, Nachrichten, ja sogar ganze Enzyklopädien sind inzwischen rund um die Uhr online verfügbar. Mithilfe von Suchmaschinen können wir diese Informationen abfragen. Allerdings stößt man gelegentlich an die Grenzen dieser Technologie. Nehmen wir beispielsweise an, wir möchten wissen, welche bekannten Wissenschaftler auch politisch aktiv sind. Diese Frage lässt sich kaum so formulieren, dass sie von Google sinnvoll beantwortet werden kann. Alle Anfragen nach *„Wissenschaftler Politiker"* geben lediglich Stellungnahmen zu politischen Ereignissen zurück. Die Ursache für dieses Problem ist, dass die Computer unserer Zeit zwar Unmengen an Daten speichern, aber weit davon entfernt sind, diese in einen Kontext einzuordnen oder gar zu *„verstehen"*. Wenn es gelänge, dem Computer diese Daten als *„Wissen"* begreiflich zu machen, so könnte dieses Wissen nicht nur bei der Internetsuche helfen, sondern auch bei vielen anderen Aufgaben wie beispielsweise bei der automatischen Übersetzung eines Textes in mehrere Sprachen oder beim Verstehen gesprochener Sprache. Dies ist das Ziel des Projektes „YAGO-NAGA" am Max-Planck-Institut für Informatik.

Damit der Computer das Wissen überhaupt verarbeiten kann, muss es auf eine strukturierte Art abgespeichert werden. Eine solche strukturierte Wissensammlung heißt *„Ontologie"*. Die Bausteine einer Ontologie sind *„Entitäten"*. Eine Entität ist jede Art von konkretem oder abstraktem Objekt: Der Physiker Albert Einstein, das Jahr 1879 oder der Nobelpreis. Die Entitäten sind durch *„Relationen"* miteinander verbunden. So ist beispielsweise Albert Einstein über die Relation *„geboren"* mit dem Jahr 1879 verbunden (siehe Grafik). Wir haben nun einen Ansatz entwickelt, der so eine Ontologie automatisch erstellt. Dazu nutzen wir die Online-Enzyklopädie Wikipedia. Wikipedia enthält Artikel zu Abertausenden von Persönlichkeiten, Produkten und Organisationen. Jeder dieser Artikel wird eine Entität in unserer Ontologie. Zum Beispiel gibt es einen Artikel über Albert Einstein, sodass der Physiker als Entität in die Ontologie aufgenommen werden kann. Jeder Artikel in Wikipedia ist bestimmten Kategorien zugeordnet. So befindet sich beispielsweise der Artikel über Einstein in der Kategorie *„Gestorben 1955"*. Dadurch kann der Computer den Fakt aufnehmen, dass Einstein 1955 gestorben ist. Dies funktioniert einfach auf Basis der Schlüsselwörter in den Kategoriebezeichnungen, ohne dass der Computer dazu den vollen Text des Artikels verstehen müsste. Dadurch erhalten wir eine sehr große Ontologie, in der alle in Wikipedia bekannten Entitäten ihren Platz haben. Diese Ontologie heißt YAGO (Yet Another Great Ontology, http://www.mpi-inf.mpg.de/yago-naga/yago/). Momentan (September 2009) enthält YAGO 2 Millionen Entitäten und rund 20 Millionen Fakten.

Dies ist schon eine beträchtliche Menge an Allgemeinwissen. Sie dient nun als Ausgangspunkt für das weitere Sammeln von Wissen. Das Ziel ist es, auch andere Internetdokumente wie Biographien, Lexikoneinträge, Homepages und Nachrichtentexte inhaltlich für den Computer zu erschließen. Diese Aufgabe ist nicht einfach. Nehmen wir beispielsweise an, der Computer fände in einer Biografie den Satz *„Einstein wurde 1879 geboren"*. Dieser Satz ist für den Computer lediglich eine Folge von Buchstaben. Für den Computer sieht der Satz also so aus, wie für die meisten von uns die chinesische Zeichenfolge.

爱因斯坦出生於1879年

Wenn der Computer aber bereits Einstein und das Datum 1879 in der Ontologie verzeichnet hat, so kann er den Satz sehr viel besser analysieren. Wir haben daher ein Verfahren entwickelt, welches das bereits vorhandene Wissen in YAGO dazu ausnutzt, neues Wissen zu finden und zu der Ontologie hinzuzufügen. Dieses Verfahren haben wir SOFIE genannt (*Self-organizing Framework for Information Extraction*, http://www.mpi-inf.mpg.de/yago-naga/sofie/). In der Tat kennt YAGO heutzutage mehrere Hundert Personen, die sowohl Wissenschaftler als auch Politiker sind. In unserem nächsten Artikel beschreiben wir, wie man das Wissen in YAGO abfragen kann. :::



KONTAKT

Gjergji Kasneci
ABT. 5 Datenbanken und Informationssysteme
Telefon +49 681 9325-500
Email kasneci@mpi-inf.mpg.de

Fabian Suchanek
ABT. 5 Datenbanken und Informationssysteme
Telefon +49 681 9325-500
Email suchanek@mpi-inf.mpg.de

Gerhard Weikum
ABT. 5 Datenbanken und Informationssysteme
Telefon +49 681 9325-500
Email weikum@mpi-inf.mpg.de
Internet http://www.mpi-inf.mpg.de/yago-naga/yago/
http://www.mpi-inf.mpg.de/yago-naga/sofie/

Die NAGA-Engine zur Suche nach Wissen statt nach Webseiten

Eine Ontologie wie YAGO ist nur von begrenztem Nutzen, wenn man sie nicht abfragen kann. Deshalb haben wir NAGA (Not Another Google Answer) entwickelt. NAGA ist eine Suchmaschine, die die explizite Struktur der YAGO-Daten versteht und ausnutzt. Sie ermöglicht Fragen nach genauem Wissen, wie zum Beispiel:

- *Welche Politiker sind auch Wissenschaftler?*
- *Welche Studenten von Max Planck haben auch einen Nobelpreis bekommen?*
- *Welche Beziehungen gibt es zwischen Niels Bohr, David Bohm, Richard Feynman und Enrico Fermi?*

Heutige Suchmaschinen wie Google, Yahoo oder Bing können mit Fragen dieser Art nicht umgehen. Zum einen könnten die Antworten auf mehreren Webseiten verstreut sein. Des Weiteren können die Suchmaschinen die Bedeutung der Frage nicht verstehen. Zum Beispiel verstehen sie nicht, dass die Wörter *„Wissenschaftler"* und *„Politiker"*, aus der ersten Anfrage in unserem Suchkontext, eine besondere Bedeutung haben. Das heißt wir sind nicht einfach an Webseiten interessiert, die beide Worte enthalten, sondern an Personen, die sowohl Wissenschaftler als auch Politiker sind.

NAGA verwendet eine Anfragesprache, die auf die graphbasierte Struktur der YAGO-Daten zugeschnitten ist. Dies erlaubt NAGA, zum Beispiel, Personen zu finden, die in YAGO sowohl als Politiker als auch als Wissenschaftler vorkommen. Die beiden Abbildungen veranschaulichen die graphbasierte Struktur von NAGA-Anfragen. Die Anfrage in Abbildung 1 fragt nach Personen, die sowohl Politiker als auch Wissenschaftler sind. Die Kantenbeschriftungen stehen für Relationen und die Knotenbeschriftungen für Entitäten. Das $x-Zeichen hat die Funktion eines Platzhalters. Die Anfrage in Abbildung 2 verwendet den regulären Ausdruck .* an den Kanten, um nach Beziehungen zwischen Niels

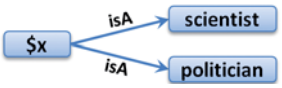


Abbildung 1

Abbildung 2

Bohr, Enrico Fermi und Richard Feynman zu fragen. Nun kann so eine Anfrage leicht mehrere Dutzend oder gar mehrere Tausend Ergebnisse liefern. Wie in der Internet-Suche auch ist es daher unumgänglich, die Ergebnisse so zu sortieren, dass die wichtigen Ergebnisse zuerst erscheinen. Im obigen Beispiel möchte der Benutzer wohl Persönlichkeiten wie Benjamin Franklin oder Angela Merkel an erster Stelle sehen.

Um die Wichtigkeit eines Ergebnisses im Bezug auf die Anfrage zu ermitteln, verwendet NAGA ein statistisches Verfahren, das die Redundanz und Vielfalt der Information im Web ausnutzt. Wenn wir zum Beispiel nach Personen suchen, die Physiker sind, und Albert Einstein und Max Mustermann zwei mögliche Antworten sind, dann ermittelt NAGA die Wichtigkeit der Antwort basierend auf der Häufigkeit, mit der Albert Einstein im Web als Physiker auftaucht und der Häufigkeit, mit der Max Mustermann im Web als Physiker auftaucht. Da Albert Einstein eine berühmte Persönlichkeit unter den Physikern ist, wird die Anzahl der Webseiten, die über ihn als Physiker sprechen, höher sein. Deswegen wird Einstein in NAGAs Ergebnisliste einen höheren Rang als Max Mustermann haben.

NAGA erlaubt auch das Auffinden von Beziehungen zwischen Entitäten. Zum Beispiel können wir erfragen, über welche Fakten die vier Physiker Niels Bohr, David Bohm, Richard Feynman und Enrico Fermi miteinander verbunden sind. Wegen der großen Menge an Fakten in YAGO war eine solche Anfrage vormals kaum effizient zu beantworten. Daher haben wir einen Algorithmus namens STAR *(Steiner Tree Approximation in Relationship Graphs)* entwickelt. STAR nutzt die Struktur von Ontologien wie YAGO aus, um Beziehungen der obigen Art effizient ausfindig zu machen. In einem ersten Schritt findet STAR eine taxonomische Beziehung zwischen den Anfrageentitäten. Eine taxonomische Beziehung zwischen Niels Bohr, David Bohm, Richard Feynman und Enrico Fermi wäre, dass alle vier Entitäten Wissenschaftler sind. Diese Beziehung wird dann nach und nach zu kompakteren und interessanteren Beziehungen verbessert. So können wir zum Beispiel herausfinden, dass alle vier Entitäten Quanten-Physiker sind, und dass alle vier am Manhattan-Projekt teilgenommen haben. NAGA kann online auf http://www.mpii.de/yago-naga/naga/ ausprobiert werden. :::

KONTAKT

Gjergji Kasneci
ABT. 5 Datenbanken und Informationssysteme
Telefon +49 681 9325-500
Email kasneci@mpi-inf.mpg.de

Fabian Suchanek
ABT. 5 Datenbanken und Informationssysteme
Telefon +49 681 9325-500
Email suchanek@mpi-inf.mpg.de

Gerhard Weikum
ABT. 5 Datenbanken und Informationssysteme
Telefon +49 681 9325-500
Email weikum@mpi-inf.mpg.de
Internet http://www.mpii.de/yago-naga/naga/

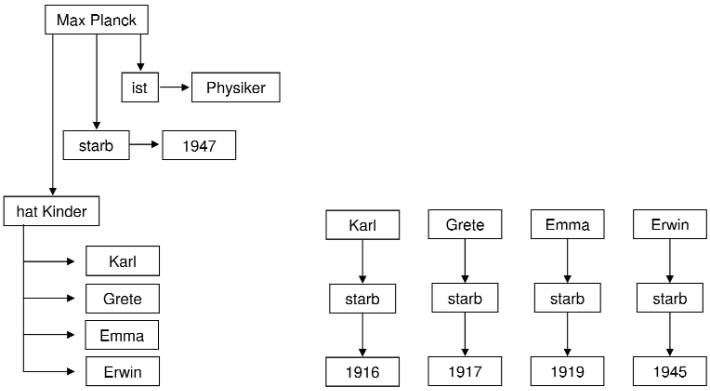
Entscheidungsverfahren für Ontologien

Die Antworten auf viele Fragen lassen sich bereits heute mit Hilfe von Suchmaschinen im Internet finden. Allerdings erlauben diese Suchmaschinen eine Suche nur rein syntaktisch anhand von Schlüsselwörtern. So extrahiert ein Suchdienst die Schlüsselwörter *„Physiker“*, *„Kind“* und *„lebte“* aus der Anfrage *„Welcher Physiker lebte länger als alle seine Kinder“* und liefert als Ergebnis Dokumente, die diese Schlüsselwörter enthalten. Die Wörter der Anfrage müssen in den gefundenen Dokumenten exakt vorkommen. Auf die obige Anfrage bekommt man eine Fülle von Dokumenten, die die Wörter *„Physiker“*, *„Kind“* und *„lebte“* enthalten, aber typischerweise nicht das gesuchte Ergebnis. Eine Umformulierung der Frage kann zum gewünschten Ergebnis führen – in unserem Fall: *„Welcher Physiker überlebte alle seine Kinder“*. Diese Anfrage liefert mit hoher Wahrscheinlichkeit als Treffer eine Internetseite, die die gesuchte Information enthält. Das Problem liegt hier darin, dass die Suchmaschine die Bedeutung der Informationen in den Dokumenten und der Anfrage nicht kennt, sondern nur rein syntaktisch mit den Wörtern, d.h. Zeichenketten, arbeitet.

Damit die Bedeutung von Wörtern und Sätzen und das damit transportierte Wissen von einem Computer verarbeitet und verstanden werden kann, muss dieses Wissen in einer Struktur abgespeichert werden. Solch eine Struktur nennt man Ontologie. Für bestimmte in der Praxis häufig auftretenden Fragestellungen wurden bereits, auf Ontologien basierende Suchverfahren entwickelt. Die Artikel *„YAGO – eine digitale Wissensammlung“*, Seite 58 und *„Die NAGA-Engine zur Suche nach Wissen statt nach Webseiten“*, Seite 59 bieten eine detaillierte Beschreibung dieser Verfahren.

Die Herausforderung des gegenwärtigen Projekts besteht darin, nicht nur Antworten auf einige bestimmte Fragestellungen zu bekommen, sondern alle Fragen zu beantworten, deren Antwort sich aus dem Wissen einer vorliegenden Ontologie folgern lässt. Dazu ist es zunächst notwendig die Ontologie in eine für das allgemeine Schließen geeignete Sprache, eine so genannte beschreibende Logik, zu überführen. Eine in Logik übersetzte Ontologie hat die Eigenschaft, dass alles hergeleitet werden kann, was aus der Ontologie folgt. Um jetzt effektiv Fragen beantworten zu können, muss die Ontologie saturiert werden, d.h. sie wird in eine kompakte Darstellung aller logischen Konsequenzen überführt. Mit Hilfe dieser Darstellung kann man dann effizient Antworten auf Fragen finden, die aus der ursprünglichen Ontologie folgen. Allerdings sind die zur Saturierung notwendigen Operationen auf der Logik sehr rechenintensiv, was die üblichen Saturationsverfahren für Ontologien mit mehreren 10 Millionen Einträgen unbrauchbar macht. Die Operationen auf die Struktur von Ontologien so abzustimmen, dass eine effiziente Saturation möglich ist, ist die Herausforderung, die es zu bewältigen gilt, um effektive Entscheidungsverfahren zu bekommen.

Ein weiteres Ziel ist es, Ontologien automatisch mit Wissen aus dem Internet zu erweitern. Dabei stellt sich aber das Problem, dass sich an verschiedenen Stellen im Internet sich widersprechende Informationen befinden. Hat man widersprüchliche Informationen in der Ontologie und löst diese nicht auf, so kollabiert die Ontologie wegen ex falso quodlibet und das bereits gesammelte Wissen wird unbrauchbar. Der Mensch ist in der Lage Informationen aus verschiedenen Quellen zu differenzieren und somit die Konsistenz sicher zu stellen. Für einen Computer ist dies allerdings keine leichte Aufgabe. Hier ermöglicht das Überführen in die Logik dem Computer die gegensätzlichen Informationen zu erkennen und dann zu differenzieren. Mit Hilfe von Konfidenzwerten lassen sich die verschiedenen Gewichtungen von gesammeltem Wissen modellieren und somit Konflikte auflösen. :::



Frage: Welcher Physiker lebte länger als alle seine Kinder?
Antwort: Max Planck

KONTAKT

Christoph Weidenbach
FG. 1 Automatisierung der Logik
Telefon +49 681 9325-900
Email weidenbach@mpi-inf.mpg.de

Patrick Wischnewski
FG. 1 Automatisierung der Logik
Telefon +49 681 9325-908
Email wischnew@mpi-inf.mpg.de

OPTIMIERUNG

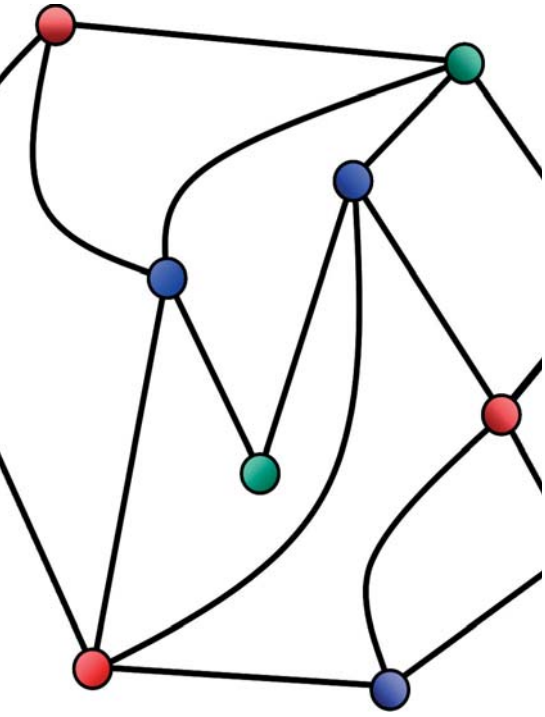
Optimierungsverfahren sind heutzutage von zentraler Bedeutung für die Effektivität von Unternehmen. Sie werden zum Beispiel eingesetzt, um den Bedarf an teuren Ressourcen wie etwa Arbeit oder Rohstoffen einzusparen. Die Herausforderung an die Wissenschaft ist es, effektive Verfahren zum Lösen von Optimierungsproblemen zu entwickeln. Mit Hilfe dieser Verfahren sollen sich schnell optimale Lösungen finden lassen oder zumindest solche, die nahe am Optimum liegen.

Gute Optimierungsverfahren sind in verschiedensten Bereichen von zentraler Bedeutung. Für große Unternehmen sind sie zum Beispiel entscheidend für ihre Wettbewerbsfähigkeit. Durch sorgfältige Planung können in Industrieprojekten oft große Mengen an Ressourcen eingespart werden, was zu geringeren Kosten führt. Allerdings sind solche Planungsprobleme oft sehr komplex und haben viele unterschiedliche Anforderungen zu berücksichtigen.Das macht es für den Computer schwer, optimale oder zumindest sehr gute Lösungen zu finden.

Am Max-Planck-Institut für Informatik beschäftigen wir uns mit derartigen schwierigen Optimierungsproblemen aus verschiedensten Anwendungsbereichen wie der industriellen Optimierung oder der Medizin. Zum einen entwickeln wir ausgefeilte Verfahren, um höchst effizient optimale Lösungen zu finden. Ist das zugrunde liegende Problem zu schwierig, um eine optimale Lösung schnell zu berechnen, entwickeln wir Verfahren, um zumindest eine Lösung zu finden, die nahe am Optimum liegt. Außerdem erforschen wir in welcher Weise die Verwendung von Zufallsentscheidungen zu effektiveren und einfacheren Optimierungsverfahren führen kann. Hierbei betrachten wir auch Verfahren, welche durch Optimierungsprozesse in der Natur inspiriert sind. Solche Verfahren ermöglichen es oft eine gute Lösung für ein gegebenes Problem ohne viel Entwicklungsaufwand zu erzielen.

Da die Optimierung in sehr vielen Bereichen eine wesentliche Rolle spielt, untersuchen Wissenschaftler aus allen Forschungsgebieten, die am Max-Planck-Institut betrachtet werden, Optimierungsprobleme. Optimierung ist heutzutage ein wesentlicher Schlüssel zur Sicherstellung effizienter Abläufe. Diese Bedeutung wird auch in Zukunft weiter zunehmen. :::

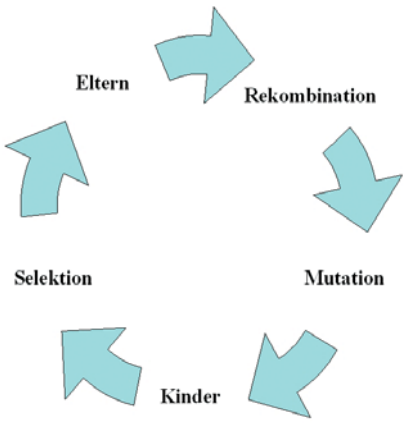
BEITRÄGE



Theorie evolutionärer Algorithmen

Evolutionäre Algorithmen *(EAs)* sind allgemeine Suchverfahren, die in den Ingenieurdisziplinen und im Bereich der kombinatorischen Optimierung vielfältig angewendet werden. Diese Klasse von Lösungsverfahren folgt dem Vorbild der Evolution und dem Darwinschen Prinzip des *„survival of the fittest"*. Angelehnt an das natürliche Evolutionsprinzip, wird ein spezieller Lösungskandidat als *Individuum* und eine Menge solcher Kandidaten als *Population* bezeichnet. Eine so genannte Fitnessfunktion, welche vom gegebenen Problem abhängt, bewertet die Lösungskandidaten. Nach dem biologischen Prinzip wird aus einer Eltern-Population eine Kinder-Population erzeugt. Dies geschieht durch so genannte Veränderungsoperatoren, die das genetische Material der Eltern an die Kinder vererben. Die wichtigsten Operatoren sind in diesem Fall *Rekombination* und *Mutation*. Die Rekombination erzeugt gewöhnlich aus zwei Eltern ein Kind, während die Mutation zusätzlich dafür sorgt, dass das Kind weitere neue Eigenschaften aufweist. Von einer Startpopulation ausgehend, ist es das Ziel, für ein gegebenes Problem eine Menge möglichst guter Lösungskandidaten zu erhalten. Nachdem zunächst durch Veränderungsoperationen Kinder erzeugt worden sind, werden anhand der Fitnessfunktion aus der Eltern-Kind-Menge Individuen ausgesucht und eine so neue Elternpopulation geschaffen.

Evolutionäre Algorithmen werden insbesondere dann eingesetzt, wenn für ein gegebenes (neues) Problem kein guter problemspezifischer Algorithmus vorhanden ist. Es kann nicht erwartet werden, dass EAs speziell für ein Problem entworfene Lösungsverfahren übertreffen. Es ist also nicht Ziel der Forschung, zu zeigen, dass EAs problemspezifischen Algorithmen überlegen sind. Vielmehr steht im Vordergrund, die Arbeitsweise evolutionärer Verfahren zu verstehen.



Ablaufschema eines evolutionären Algorithmus

Forschungsschwerpunkt

Während evolutionäre Verfahren bereits vielfach erfolgreich angewendet werden, steckt das theoretische Verständnis dieser Algorithmen im Vergleich zu klassischen Algorithmen noch in den Kinderschuhen.

Wir untersuchen, wie evolutionäre Suchverfahren in der Lage sind, bestimmte Probleme zu lösen, und mit welchen Strukturen EAs gut oder schlecht umgehen können. Das Hauptaugenmerk ist darauf gerichtet, wie viel Zeit EAs benötigen, um für ein gegebenes Problem eine optimale Lösung zu generieren. Da evolutionäre Algorithmen eine spezielle Klasse randomisierter Algorithmen sind, kann man auf eine große Zahl klassischer Analysemethoden zurückgreifen. Des Weiteren werden neue Analysemethoden entwickelt, die insbesondere evolutionäre Verfahren analysieren.

Es zeigt sich, dass evolutionäre Algorithmen oft gute Lösungen für bekannte Probleme finden. Sie sind bei vielen Problemen in der Lage, sich ähnlich wie problemspezifische Algorithmen zu verhalten. So wurde gezeigt, dass evolutionäre Algorithmen kürzeste Wege zwischen allen Knoten in einem gegebenen Graphen effizient berechnen können. Die durchgeführten Analysen zeigen, dass die Verwendung von Rekombination und Mutation einen beweisbaren Vorteil gegenüber evolutionären Algorithmen bringen, welche lediglich Mutation als Variationsoperator verwenden.

Andere Studien zeigen, dass Ansätze der multikriteriellen Optimierung evolutionären Algorithmen zusätzliche Möglichkeiten der effizienten Suche geben. Viele Optimierungsprobleme sind durch eine Zielfunktion gegeben, welche unter eine Menge von Nebenbedingungen optimiert werden soll. In dem multikriteriellen Ansatz werden nun diese Nebenbedingungen als zusätzliche gleichwertige Zielfunktionen betrachtet. Dieses gibt der Suche evolutionärer Algorithmen zusätzliche Suchrichtungen, was bei verschiedenen kombinatorischen Optimierungsproblemen zu effizienteren Verfahren führt. :::

KONTAKT

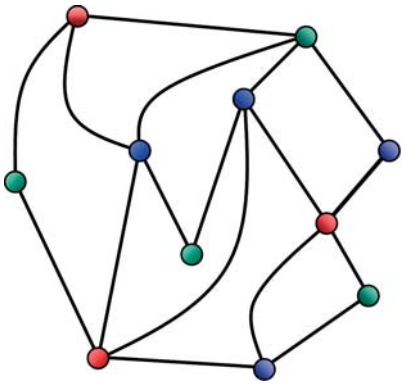
Benjamin Doerr
ABT. 1 Algorithmen und Komplexität
Telefon +49 681 9325-104
Email doerr@mpi-inf.mpg.de

Frank Neumann
ABT. 1 Algorithmen und Komplexität
Telefon +49 681 9325-117
Email fne@mpi-inf.mpg.de

Zufällige Strukturen in der Informatik

Erfüllbarkeitsprobleme

Erfüllbarkeitsprobleme mit Nebenbedingungen gehören zu den wichtigsten Problemen in der theoretischen Informatik und haben zahlreiche Anwendungen in praktischen Fragestellungen. Ein typisches Erfüllbarkeitsproblem besteht aus n Variablen, denen Werte aus einer bestimmten Menge zugeordnet werden müssen. Zusätzlich sind *Bedingungen* gegeben, die jeweils einige der Variablen binden, also die Werte, die zu diesen Variablen zugeordnet werden dürfen, in gewisser Weise einschränken. Ein klassisches Beispiel für ein Erfüllbarkeitsproblem ist das Färben von Graphen. Gegeben ist ein Graph, also eine Menge von Knoten die durch Kanten verbunden sind, und eine Zahl k. Die Frage lautet: Lässt sich jedem Knoten eine dieser Farben zuordnen, so dass alle Kanten unterschiedlich gefärbte Endpunkte haben? Dieses abstrakte Problem hat signifikante Anwendungen in vielen unterschiedlichen Bereichen, wie zum Beispiel die Erstellung von Fahrplänen der Deutschen Bahn oder die Zuordnung von Frequenzbändern in mobilen Netzwerken.



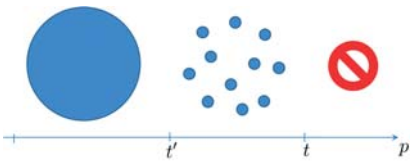
Ein Graph der mit drei Farben zulässig färbbar ist

Das Färbungsproblem sowie fast sämtliche Erfüllbarkeitsprobleme lassen sich offensichtlicherweise durch einen einfachen Ansatz lösen: Man probiere für jede Variable alle möglichen Zuordnungen aus, entweder bis man eine zulässige Belegung gefunden hat, oder überzeugt ist, dass es keine gibt. Dieses Verfahren benötigt allerdings exponentielle Laufzeit: Im Fall des Färbungsproblems sind es im schlimmsten Fall mindestens k^n Schritte, falls der zugrundeliegende Graph n Knoten hat. Eines der grundlegendsten offenen Probleme der theoretischen Informatik ist es zu verstehen, ob es möglich ist *signifikant* schnellere Verfahren als die vollständige Aufzählung aller möglichen Lösungen zu entwickeln. Solche Verfahren heißen effiziente Algorithmen und benötigen nur polynomiell viele Schritte als Funktion der Eingabegröße. Unglücklicherweise hat die Theorie in den letzten 50 Jahren keine bahnbrechenden Fortschritte bezüglich dieser Fragestellung gemacht.

Zufällige Instanzen

Um die Wurzeln des Problems zu verstehen haben Informatiker, Mathematiker und Physiker in den letzten 25 Jahren einen möglichen Zusammenhang zwischen effizienter Berechenbarkeit und Phasenübergängen in zufälligen Instanzen von Erfüllbarkeitsproblemen untersucht. Im Falle des Färbungsproblems konstruiert man eine zufällige Instanz indem man unabhängig für jedes Paar von Knoten mit Wahrscheinlichkeit p eine Kante in einen anfangs leeren Graphen hinzufügt. Es ist bekannt, dass ein kritisches t existiert, so dass falls $p < t$, dann ist der so entstehende Graph mit Wahrscheinlichkeit sehr nahe an 1 mit k Farben zulässig färbbar, und falls $p > t$, so ist er es nicht. Andererseits, aus einem algorithmischen Blickwinkel betrachtet, versagen alle bekannten effizienten Algorithmen schon bei Graphen mit Kantenwahrscheinlichkeiten die signifikant kleiner sind als t. Genauer, es wurde beobachtet, dass es ein $t' < t$ gibt, so dass die besten Algorithmen eine zulässige Färbung finden nur falls $p < t'$, und sonst nicht.

Im Jahr 2008 wurde bewiesen, dass diese Beobachtung nicht nur reine Spekulation ist. Der Punkt t', an dem alle Algorithmen versagen, fällt mit dem Punkt zusammen, an dem der Raum der zulässigen Färbungen eines zufälligen Graphen sich dramatisch verändert. An dieser Stelle präsentieren wir nur eine stark vereinfachte Version des tatsächlichen Phänomens. Falls $p < t'$, so sieht die Menge der zulässigen Färbungen aus wie ein gigantischer Ball: Man kann jede zulässige Färbung erhalten, indem man irgendwo anfängt und die Farben einzelner Knoten schrittweise verändert. Im Gegensatz dazu, am Punkt $p = t'$ zerspringt dieser Ball in exponentiell viele kleine Stücke, die sehr weit voreinander weg sind. Viele Algorithmen können im ersten Regime leicht zulässige Färbungen finden, aber es ist kein Algorithmus bekannt, der im zweiten Regime erfolgreich ist.



Die Struktur der zulässigen Färbungen eines Zufallsgraphen

In unserer Arbeitsgruppe beschäftigen wir uns intensiv mit der Analyse der Struktur von zufälligen Instanzen kombinatorischer Erfüllbarkeitsprobleme. Unser Ziel ist es die relevanten Eigenschaften zu entdecken und zu verstehen, die die Leistung von effizienten Algorithmen und deren Entwurf maßgeblich beeinflussen. Zusätzlich entwickeln wir notwendige mathematische Hilfsmittel, die uns ermöglichen präzise Aussagen über die zugrundeliegenden Phänomene zu machen. :::

KONTAKT

Konstantinos Panagiotou
ABT. 1 Algorithmen und Komplexität
Telefon +49 681 9325-113
Email kpanagio@mpi-inf.mpg.de

Planen unter Unsicherheit

Beim Lösen von Optimierungsaufgaben in der Praxis sind unvollständige und unsichere Daten ein allgegenwärtiges Problem. So plant ein Pizzaservice beispielsweise seine Liefertouren ohne zukünftige Bestellungen zu kennen; die Belegung von Computerprozessoren wird geplant, obwohl lediglich stochastische Abschätzungen der individuellen Dauern bekannt sind; Produktionsabläufe werden zeitlich genau geplant und doch können Maschinen unerwartet ausfallen; und die Einhaltung eines Terminplans, z.B. von Bauprojekten, hängt von Unsicherheitsfaktoren wie Wetter, Krankheit, etc. ab.

In unserer Arbeitsgruppe beschäftigen wir uns mit dem Umgang mit unvollständigen Informationen beim Lösen von Optimierungsaufgaben. Wir entwikkeln Algorithmen, die gute Lösungen erzielen, obwohl Inputdaten stochastisch verteilt sind oder zum Planungszeitpunkt noch gar nicht bekannt sind. Dabei untersuchen wir insbesondere Fragestellungen aus dem Schedulingbereich. Als *Scheduling* bezeichnet man die zeitliche Zuordnung von Vorgängen auf Ressourcen mit beschränkter Kapazität. Dabei wird ein Optimierungsziel verfolgt, wie z.B. die Minimierung der Gesamtprojektdauer, die Maximierung des Durchsatzes oder die Minimierung von Kundenwartezeiten.

Online-Scheduling

Im Online-Scheduling betrachten wir Planungsaufgaben, bei denen eine Probleminstanz erst Stück für Stück während des Planungsprozesses bekannt wird; das heißt, zu jedem Zeitpunkt müssen Entscheidungen getroffen werden, die lediglich auf den bisher bekannten Daten beruhen.

Ein Pizzaservice beispielsweise erhält seine Bestellungen telefonisch über den Abend verteilt. Wenn er alle Bestellungen und deren zeitliche Eingänge genau im Voraus kennen würde, könnte er einen optimalen Zeitplan für die Auslieferungstouren aufstellen. In der Realität handelt es sich jedoch um ein Online-Problem, bei dem die Bestellungen eben nicht im Voraus bekannt sind. Sobald ein Lieferauftrag eingeht, muss der Planer die Entscheidung balancieren zwischen einer zeitnahen, kundenfreundlichen Lieferung und dem Warten auf potentielle weitere kurzfristige Lieferaufträge in die gleiche Gegend die sich kostengünstig kombinieren lassen.

Für solche Planungsaufgaben entwickeln wir Lösungsalgorithmen, die trotz unvollständiger Information, gute Pläne aufstellen. Typischerweise bewertet man die Güte dieser Algorithmen im Worst-Case-Vergleich mit einer bestmöglichen offline Lösung – man nennt dies *Kompetitive Analyse*. Das heißt, man vergleicht den Plan, den ein Algorithmus unter Unsicherheit erstellt, mit dem optimalen Plan eines allwissenden Planers, der alle Daten, z.B. die Bestelleingänge, im Voraus kennt. Für viele Problemklassen führt die Kompetitive Analyse zu einer klaren Unterscheidung und sinnvollen Bewertung von Algorithmen. In einigen Fällen jedoch ist sie zu pessimistisch, d.h. der Vergleich gegen einen allwissenden Planer ist unfair und lässt jeden Online-Algorithmus gleich schlecht aussehen. Daher beschäftigen wir uns auch mit alternativen Ansätzen zur Bewertung von Online-Algorithmen.

Stochastisches Scheduling

Eine andere Art von Unsicherheit wird im stochastischen Scheduling berücksichtigt. Hier werden die Dauern der zu planenden Vorgänge als Zufallsvariablen modelliert. Zu Beginn der Planung sind alle Vorgänge mit ihren problemspezifischen, deterministischen Daten und den Verteilungsfunktionen für die Vorgangsdauern bekannt. In der Praxis können solche Informationen aus statistischen Auswertungen oder Erfahrungswerten gewonnen werden. Die tatsächliche Realisierung der Dauer erfährt der Planer erst während der Bearbeitung eines Vorgangs.

Ziel ist es eine Schedulingpolitik zu entwickeln, die in Erwartung einen guten Schedule findet. Formal gesprochen suchen wir Politiken, deren erwarteter Zielfunktionswert im Worst-Case nur um einen konstanten, instanzunabhängigen Faktor von dem Erwartungswert einer optimalen stochastischen Politik abweicht. Man beachte dass dies ein fairer Vergleich ist gegen eine bestmögliche Politik, die auch gegen die Unsicherheit bestehen muss.

Für dieses Modell zum Planen unter Unsicherheit sind kaum theoretische Resultate bekannt. Als besondere Hürde stellt sich das Finden von unteren Schranken an den zu erwartenden Zielfunktionswert optimaler Politiken dar. Hier konnten wir Techniken entwickeln, die zu neuen unteren Schranken führen. Damit können wir für ein spezielles Schedulingproblem, bei dem Vorgänge unterbrochen und später weiterbearbeitet werden dürfen, erstmalig beweisbar gute Schedulingpolitiken entwerfen. Diese Erkenntnisse sind auch in anderen Bereichen der Optimierung unter Unsicherheit von Bedeutung. :::

KONTAKT

Nicole Megow
ABT. 1 Algorithmen und Komplexität
Telefon +49 681 9325-118
Email nmegow@mpi-inf.mpg.de

Messoptimierung für medizinische Bildgebung

Optimierung von Magnet-Resonanz-Tomographie Akquisition

In der Magnet-Resonanz-Tomographie (MRT) werden Bilder aus linearen Fouriermessungen durch digitale Berechnung rekonstruiert. Moderne nichtlineare Methoden kommen mit weniger Daten aus als herkömmliche lineare Schätzer, wobei eine präzise Auswahl der Messprojektionen (des Designs) für gute Ergebnisse unerlässlich ist. Messzeit, die wesentliche Beschränkung heutiger MRT, wird durch unterabgetastete Rekonstruktion wesentlich verkürzt.

Bayessche Bewertung von Designs

Wie kann das Design mit geringem menschlichen Experten- und tomographischem Messaufwand optimiert werden? Unser Ansatz entspringt der Bayesschen Statistik: Gute Designs lösen möglichst viel Unsicherheit über das zu schätzende Bild auf. Die Optimierung erfolgt sequenziell, zerlegt in MRT-Phasenkodierschritte, und unüberwacht: In jedem Schritt werden Kandidaten durch Bayessche Berechnungen bewertet, der Gewinner dem Design hinzugefügt. Dies entspricht einer adaptiven Variante von Compressive Sensing, unterstützt durch Konzepte des maschinellen Lernens. Ohne von vornherein einschränkende Annahmen wird das Design auf realen Trainingsdaten optimiert.

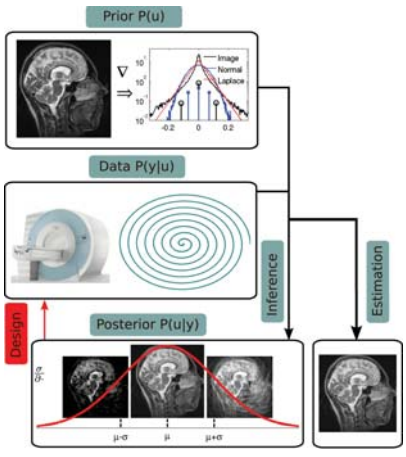
Bayessche Inferenz, also die Quantifikation von Unsicherheiten bei der Rekonstruktion, ist ungleich schwieriger zu approximieren als letztere allein. Ein neues variationelles Verfahren erlaubt uns, Inferenz in bisher nicht erreichten Größenordnungen über hochaufgelösten Bildern zu approximieren. Dies gelingt durch iterative Reduktion auf in der Bildverarbeitung gängige Teilprobleme, wobei die für uns wesentlichen Posterior-Kovarianzen direkter als in bisherigen Algorithmen erfasst werden.

Optimierte unterabgetastete kartesische MRT-Bildgebung

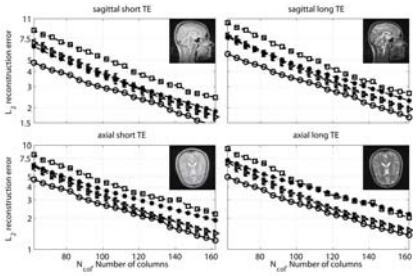
In Zusammenarbeit mit dem Max-Planck-Institut für Biologische Kybernetik, Tübingen, erfolgte eine erste Studie zur kartesischen Unterabtastung von anatomischen MRT-Bildern des menschlichen Gehirns. Wir verglichen andere vorgeschlagene Methoden zur Designwahl mit unserer Bayesschen Optimierung durch den Fehler nichtlinearer Rekonstruktion von Testbildern. Die optimierten Designs schneiden signifikant besser ab als alle anderen, vor allem bei hohen Unterabtastraten.Die Ergebnisse sind in einer MRT-Fachzeitschrift veröffentlicht.

Ausblicke

Zum Ausbau der prototypisch demonstrierten Methodik ist eine Parallelisierung auf Grafikkarten und Mehrprozessorsystemen geplant, in Zusammenarbeit mit Dr. Robert Strzodka. Reale MRT-Messungen kommen mit mehreren Bildschichten, mehreren Spulen und sogar als Zeitreihen. Messoptimierung stellt höchste, neuartige Anforderungen an parallele Hardware, denen das Max-Planck-Institut für Informatik wie sonst kaum ein Ort gewachsen ist. Zudem wird die Algorithmik verbessert werden, etwa durch Mehrgitterverfahren, unterstützt durch Visual Computing Exzellenz vor Ort. Jenseits der MRT-Problematik kann die Bayessche Methodik auf andere Visual Computing Szenarien angewendet werden, etwa zur Korrektur verwackelter Fotoaufnahmen oder zur Optimierung im Rahmen von Computational Photography. :::



Bayessche Design-Optimierung für Magnet-Resonanz-Tomographie



Rekonstruktionsfehler mit Designs verschiedener Größe $N_{\{col\}}$ *(256 für dichte Abtastung)* und Auswahl *(op: optimiert durch Bayessche Methode; rd: zufällig gezogen nach Lustig, Donoho, Pauli, Magn. Reson. Med. 85(6), 2007; ct: dichte Abtastung niedriger Frequenzen; eq: regelmäßige Abstände)*, für Testbilder unterschiedlicher Orientierung und Kontrast, gemittelt über 5 Schichten und 4 Versuchspersonen *(subjects)*.

KONTAKT

Matthias Seeger
ABT. 4 Computergrafik
Telefon +49 681 9325-452
Email mseeger@mpi-inf.mpg.de

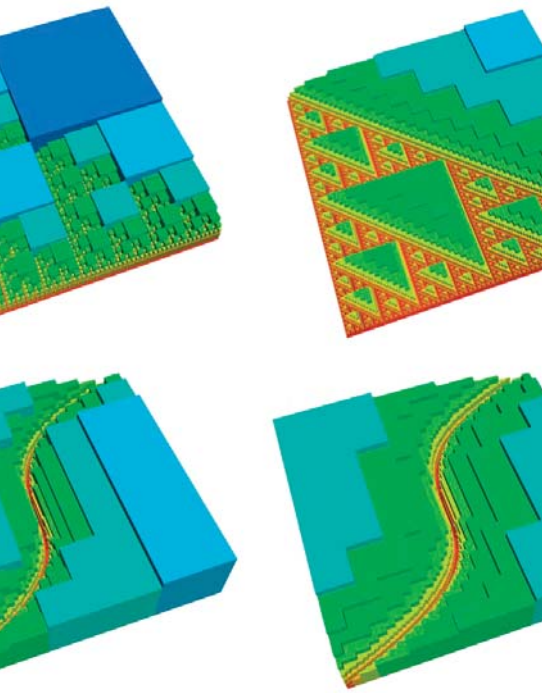
SOFTWARE

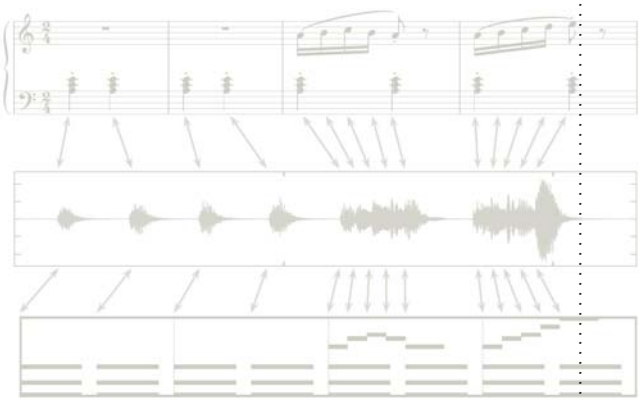
Informatik ist einerseits eine Grundlagenwissenschaft, die sich mit universellen Berechnungs- und Problemlösungsmethoden und deren fundamentalen Eigenschaften wie Korrektheit und Komplexität beschäftigt. Andererseits hat sie aber auch den Charakter einer Ingenieurwissenschaft und lebt von den vielfältigen Berührungspunkten mit verschiedensten Anwendungen. Die Grundlagenforschung trägt entscheidend zur Entwicklung von mathematischen Modellen und neuen Algorithmen für einsatzfähige Softwaresysteme bei. Programmbibliotheken und -systeme, die als Open-Source-Software mit kostenfreien Lizenzen Forschern und Anwendern zur Verfügung gestellt werden, bringen enormen Nutzen für andere Wissenschaftler und beeinflussen die langfristige Entwicklung der IT-Industrie. Nicht zuletzt liefern sie auch wichtiges Feedback für die weitere Weichenstellung der eigenen Forschungsarbeiten.

Am Max-Planck-Institut für Informatik wird diese Philosophie seit der Gründung des Instituts mit großem Erfolg verfolgt. Alle Abteilungen und Gruppen arbeiten daran, die Ergebnisse ihrer Grundlagenarbeiten in praxisrelevante Softwaresysteme umzusetzen und für Wissenschaft und Industrie verfügbar zu machen. Es gibt eine beachtliche Anzahl am Institut entwickelter Prototypsysteme aus allen Bereichen vom Theorembeweisen und der Algorithmik bis zur Bioinformatik, multimodaler Kommunikation und Web-Suche, die ihren Weg in die Wissenschaftsgemeinde gefunden haben und an vielen Orten in der Welt für Forschungsarbeiten benutzt werden. Dabei handelt es sich überwiegend um kostenfreie Open-Source-Software. In einigen Fällen wurden Startup-Firmen gegründet, die die Software weiter entwickeln und vertreiben. Beispiele für eine kommerzielle Nutzung unserer Software sind die LEDA Bibliothek für effiziente Algorithmen aus Abteilung 1, der BiQ-DNA-Methylation-Analysator aus Abteilung 3 und der Waldmeister-Gleichheitsbeweiser aus der Forschungsgruppe 1.

Die geschickte Umsetzung mathematischer Modelle und Algorithmen in lauffähige Software ist zudem selbst ein wichtiger Forschungsgegenstand. Algorithmen, die abstrakt sehr gute, mathematisch analysierbare Laufzeit- und Speicherplatzeigenschaften haben, so genannte asymptotische Komplexitätsmaße, sind in der Implementierung auf modernen Rechnern und gerade auf den aktuellen verteilten Mehrprozessor- und Cluster-Systemen nicht automatisch effizient. Eigenschaften der Prozessoren, Speicher, Magnetplatten und Kommunikation sowie die Charakteristika realer Daten müssen im Engineering geeignet berücksichtigt werden, um einsatztaugliche Softwaresysteme zu bauen. Gelingt das Softwaresystem für ein neu entwickeltes Verfahren, so liefert es wiederum wertvolle Hinweise auf relevante Spezialfälle, oder sinnvolle Generalisierungen des behandelten Problems oder der verwendeten Methoden. :::

BEITRÄGE





Zuordnung von Arbeiten an Gutachter

Konferenzen spielen in der wissenschaftlichen Kommunikation zwischen Informatikern eine zentrale Rolle. Auf Tagungen werden neueste Ergebnisse vorgetragen und diskutiert. Die Beiträge unterliegen dabei einem strengen Begutachtungsprozess. Bei den wichtigsten Tagungen werden weniger als 25 Prozent der eingereichten Arbeiten angenommen. Wegen der Bedeutung der Tagungen für den wissenschaftlichen Austausch und die Karrieren von Informatikern muss der Begutachtungsprozess möglichst effektiv organisiert werden.

Die Begutachtung wird durch ein Programmkomittee durchgeführt. Jede eingereichte Arbeit wird von mehreren Mitgliedern des Komittees begutachtet; diese dürfen dabei den Rat von anderen Wissenschaftlern einholen. Die Zuordnung der Arbeiten an die Mitglieder des Komittees wird vom Vorsitzenden vorgenommen; er benutzt dazu meist ein elektronisches Unterstützungssystem.

Kurt Mehlhorn war Vorsitzender des Programmkomittees für das European Symposium on Algorithms (ESA) 2008. 202 Arbeiten wurden zu der Tagung eingereicht und das Programmkomittee bestand aus 14 Mitgliedern. Jede Arbeit sollte von 4 Mitgliedern gesichtet werden; jedem Mitglied mussten demnach etwa $213 \cdot 4 / 14 \approx 58$ Arbeiten zugeordnet werden.

Welche Zuordnung ist vernünftig?

Wichtige Kriterien sind dabei Qualität und Fairness der Zuordnung. Wir nehmen dazu an, dass wir für jeden Gutachter und jede Arbeit eine Einschätzung kennen, die die Eignung des Gutachters für die Arbeit beschreibt, etwa *„sehr geeignet (S)“*, *„geeignet (G)“*, *„weniger geeignet (W)“* und *„nicht geeignet (N)“*. Nicht geeignet würde man immer benutzen, wenn der Gutachter einen Interessenskonflikt hat, etwa weil die Arbeit von einem seiner Studenten verfasst ist. Die folgende Tabelle zeigt ein konkretes Beispiel.

	Arbeit 1	Arbeit 2	Arbeit 3	Arbeit 4
Gutachter 1	S	S	G	W
Gutachter 2	S	W	G	G

Nehmen wir nun an, dass jede Arbeit einmal begutachtet wird und jeder Gutachter zwei Arbeiten begutachten muss. Gutachter 1 würde am liebsten die Arbeiten 1 und 2 lesen, Gutachter 2 würde am liebsten die Arbeiten 1 und 3 oder 4 lesen, für Arbeit 4 zieht Gutachter 2 vor. Das Beispiel zeigt also, dass, wie immer im Leben, nicht alle Wünsche gleichzeitig erfüllt werden können.

Sechs Zuordnungen sind möglich. Wir können die Arbeiten 1 und 2, 1 und 3, 1 und 4, 2 und 3, 2 und 4, oder 3 und 4 an Gutachter 1 zuordnen und die anderen an Gutachter 2. Welche Zuordnung sollen wir wählen? Sehen wir uns drei Beispiele an; die Zuordnung ist dabei fett hervorgehoben.

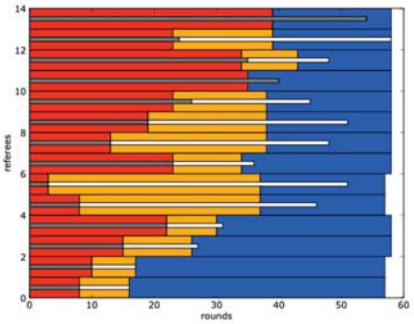
	1	2	3	4
1	S	S	**G**	**W**
2	**S**	**W**	G	G

	1	2	3	4
1	**S**	**S**	G	W
2	S	W	**G**	**G**

	1	2	3	4
1	S	**S**	**G**	W
2	**S**	W	G	**G**

Gutachter 1 zieht die mittlere Zuordnung den beiden anderen vor, Gutachter 2 zieht die dritte Zuordnung den beiden anderen vor. Die mittlere und die rechte Zuordnung haben die gleiche *globale Qualität*; beide benutzen zwei S und zwei G. Die rechte Zuordnung ist *fairer* als die mittlere Zuordnung, da sie beide Gutachter gleich behandelt. Für jeden schlägt ein S und ein G zu Buche. Die mittlere Zuordnung bevorzugt dagegen Gutachter 1.

Forschers des Max-Planck-Instituts für Informatik haben dieses Zuordnungsproblem formalisiert und effiziente Algorithmen entwickelt, die nahezu optimale Zuordnungen berechnen *(N. Garg, T. Kavitha, A. Kumar, K. Mehlhorn, J. Mestre: Assigning Papers to Referees, Algorithmica, to appear)*. Die folgenden Abbildung zeigt eine (nahezu) fairste Lösung für die ESA 2008 Instanz. Jeder Streifen entspricht einem der 14 Gutachter. Für jeden Gutachter geben die dünnen Rechtecke die Eignung an; Grau steht dabei für S und Weiß für G. Gutachter 13 (der zweite von oben) hat sich also für 25 Arbeiten als sehr geeignet und für weitere 33 als geeignet eingestuft. Für die restlichen Arbeiten war die Einstufung wenig oder nicht geeignet. Die dicken Rechtecke geben die tatsächliche Zuordnung an. Rot steht dabei für S, Ocker für G und Blau für W. Dünne Rechtecke, die über die entsprechenden dicken Rechtecken hinausgehen, stehen für unerfüllte Wünsche. Die Zuordnung erfüllt alle Wünsche von Gutachtern mit wenigen Wünschen; bei Gutachtern mit vielen Wünschen werden immer noch recht viele Wünsche erfüllt. :::



KONTAKT

Kurt Mehlhorn
ABT. 1 Algorithmen und Komplexität
Telefon +49 681 9325-100
Email mehlhorn@mpi-inf.mpg.de

GBACE: Parallele Verarbeitung adaptiver Daten

Die parallele Revolution

Etwa 2004 hat ein radikaler Umbruch bezüglich der Fabrikation von Mikroprozessoren stattgefunden. Wegen physikalischer Einschränkungen wurde die seit Jahrzehnten vorherrschende Entwicklung größerer und schnellerer Prozessorkerne aufgegeben, zugunsten einer exponentiell anwachsenden Hardware-Parallelisierung in viele Kerne innerhalb eines Chips. Diese grundlegende Veränderung im Entwurf von Mikroprozessoren stellt eine extreme Anforderung an die Software, die nun nur noch durch eine explizite Parallelisierung von dem technischen Hardware-Fortschritt profitieren kann. Automatisch wird sie auf neuer, leistungsfähigerer Hardware nicht mehr schneller.

Parallele Koprozessoren

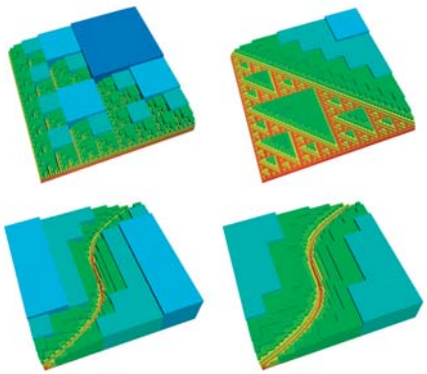
Im Jahr 2004 kamen zwar die ersten Doppelkern Prozessoren heraus, diese stellten aber noch lange nicht die gleichen Anforderungen an die Software, die erst durch die exponentiell anwachsende Zahl der Kerne mit den Jahren entstehen würden. Wir befanden uns damals erst am Anfang der Mehrkernära. Man unterscheidet grob Mehrkern *(multi-core)* Architekturen mit 2 bis 8 Kernen, Vielkern *(many-core)* Architekturen mit 16 bis 64 Kernen und massiv parallele *(massively parallel)* Architekturen mit 100 und mehr Kernen. Während die Parallelisierung auf Mehrkern-Architekturen mit traditionellen Mitteln angegangen werden kann, liegt die eigentliche Herausforderung an die Software in der Skalierung zu Vielkern und massiv parallelen Architekturen, wie Grafikprozessoren und zukünftigen Vielkern-CPUs.

Parallelität auf vielen Ebenen

Die Nutzung von Vielkern-Koprozessoren erlaubt eine starke Beschleunigung vieler Applikationen. Für größere wissenschaftliche Probleme möchte man daher gerne mehrere Koprozessoren innerhalb eines Computers verwenden, wobei das Problem auf deren Speicher Koprozessorkarten verteilt wird. Dies ist die erste Parallelisierungsebene. Eine zweite Parallelisierungsebene existiert wegen der vielen Kerne innerhalb jedes Koprozessors. In jedem Kern gibt es mehrere Recheneinheiten, die eine dritte Parallelisierungsebene formen. Für eine effiziente Software Ausführung ist es sehr wichtig, dass alle diese Parallelisierungsebenen effizient genutzt werden.

Adaptive Daten

Eine adaptive Datenrepräsentation wird oft eingesetzt, um kontinuierliche Funktionen bis zu einer gewissen Genauigkeit zu approximieren. Die Hauptidee liegt in der Benutzung einer feinen Auflösung, wo es schwierig ist die Funktion darzustellen und einer groben Auflösung, wo dies einfach ist.



Adaptive Gittererzeugung und Verfeinerung mittels paralleler Rekursion auf Grafikprozessoren

In der linken Spalte der Abbildung zeigen wir die adaptive Gitterrepräsentation zweier Funktionen. In Abhängigkeit einer Genauigkeitsschwelle haben verschiedene Regionen des Gitters verschiedene Auflösungen. In der Abbildung benutzen wir unterschiedliche Farben und Höhen der Gitterregionen für die verschiedenen Auflösungsstufen.

Für viele Anwendungen, die auf solchen adaptiven Gittern arbeiten, werden Einstufenübergänge zwischen den Auflösungen von Nachbarregionen gewünscht. Die Umsetzung dieser Kondition führt zu rekursiven Verfeinerungen vieler Regionen der adaptiven Gitter aus der linken Spalte der Abbildung. Die entsprechend verfeinerten Gitter sind in der rechten Spalte der Abbildung zu sehen. Im Vergleich zu der linken Spalte bemerken wir die sanfteren Übergänge zwischen den Auflösungen.

Parallel und adaptiv

Um wissenschaftliche Probleme effizient zu lösen, möchte man die adaptive Datenrepräsentation mit der Parallelität auf allen Ebenen kombinieren. Idealerweise geschieht dies unabhängig von dem spezifischen Vielkern-Koprozessor, könnte also mit verschiedener Hardware laufen. Es ist jedoch bereits schwierig die vielen Ebenen der Parallelität für gleich aufgelöste Probleme zu nutzen, denn mit adaptiven Daten steigt die erforderliche Softwarekomplexität wesentlich an und die hardwareunabhängige Kodierung stellt eine weitere Herausforderung dar. Wegen dieser Komplexität wird das Potenzial der Vielkern-Koprozessoren insbesondere für größere Projekte bisher wenig genutzt. Die Softwareumgebung GBACE *(General Block Adaptive Concurrent Environment)* abstrahiert die parallele Verarbeitung adaptiver Daten und berücksichtigt viele der Hardware-Erfordernisse automatisch, so dass ein einfacherer Zugang zu dem großen Potential der Vielkern-Prozessoren entsteht. Damit sollen die Vorteile der parallelen Revolution endlich auch größeren, komplizierteren Projekten zugutekommen. :::

KONTAKT

Robert Strzodka
ABT. 4 Computergrafik
Telefon +49 681 9325-427
Email strzodka@mpi-inf.mpg.de
Internet http://www.mpi-inf.mpg.de/~strzodka/software/GBACE

Automatische Erschließung von Musikdaten

Inhaltsbasierte Musiksuche

Moderne digitale Musikbibliotheken enthalten multimediale Dokumente in zahlreichen Ausprägungen und Formaten. Man denke hier beispielsweise an CD-Aufnahmen diverser Interpreten, Noten, MIDI-Daten, Musikvideos oder Gesangstexte. Allgemein gesprochen ist das Hauptziel des *Music Information Retrieval* (MIR) die Nutzbarmachung solch multimodaler und komplexer Musikdatenbestände. Eine zentrale Aufgabe ist hierbei die Entwicklung effizienter Such- und Navigationssysteme, die es dem Benutzer erlauben, den Datenbestand bezüglich unterschiedlichster musikrelevanter Aspekte zu durchsuchen. Während die textbasierte Suche nach Musik anhand von Komponistennamen, Songtitel, Werkverzeichnisnummer oder dergleichen mit klassischen Datenbanktechniken möglich ist, stellt die *inhaltsbasierte Suche* in Musikdaten ohne das Zurückgreifen auf manuell erzeugte Annotationen ein schwieriges Problem dar. Was ist zu tun, wenn man nur ein Melodiefragment vorpfeifen kann oder nur einen kurzen akustischen Ausschnitt von einem Musikstück vorliegen hat? Wie geht man vor, wenn der Benutzer an allen CD-Aufnahmen (samt der genauen Zeitpositionen innerhalb der jeweiligen Aufnahmen) interessiert ist, die gewisse Notenkonstellationen, Harmonieverläufe, oder Rhythmen aufweisen? Wie können Partiturdaten oder Musikaufnahmen hinsichtlich wiederkehrender Muster durchsucht werden? Dies ist nur eine kleine Auswahl aktueller MIR-Fragestellungen.

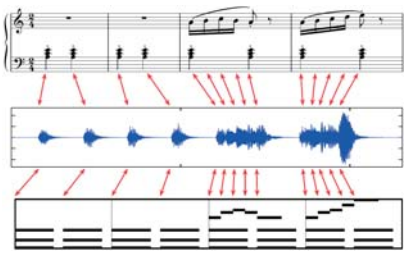


Abbildung 1: Verlinkung von Musikdaten in unterschiedlichen Formaten *(Partitur, Audio, MIDI)*, die dasselbe Musikstück *(die ersten vier Takte der Etüde Nr. 2, op. 100, F. Burgmüller)* repräsentieren.

Multimodale Verlinkung

Bei der Entwicklung inhaltsbasierter Such- und Navigationsmechanismen führt die oben angesprochene Multimodalität und Komplexität existierender Musikdokumentensammlungen zu großen, weitgehend noch ungelösten Problemen. Eine entscheidende Rolle kommt hier der umfassenden Annotation, Verlinkung und Strukturierung des Datenbestandes zu, was allerdings aufgrund der enormen Datenmassen manuell nicht bewerkstelligt werden kann. Genau diesem Punkt widmet sich die *automatisierte Musikdatenerschließung*, bei der es allgemein gesprochen um die automatische Generierung semantisch hochwertiger Annotationen geht, mittels derer dann inhaltsbasierte Anfragen an Musikdatenbanken effizient bearbeitet werden können. Grundlage bilden sogenannte Synchronisationstechniken, die zur automatischen Verlinkung zweier Datenströme unterschiedlicher Formate eingesetzt werden können. Anschaulich können solche Verfahren zu einer bestimmten Position innerhalb einer Darstellung eines Musikstücks (z.B. in einer CD-Aufnahme) die entsprechende Stelle innerhalb einer anderen Darstellung (z. B. in einer Partitur) bestimmen. [Abbildung 1]. Solche Verlinkungsdaten können dann zur multimodalen Musiknavigation und zum Vergleich unterschiedlicher Interpretation eingesetzt werden. [Abbildung 2]

Abbildung 2: Benutzerschnittstelle zum Vergleich mehrerer Interpretationen eines Musikstücks. In diesem Beispiel kann der Benutzer über die jeweiligen Slider zeitgleich in fünf unterschiedlichen Aufnahmen von Beethovens fünfter Symphonie navigieren.

Audiomatching und Strukturanalyse

Insbesondere die Analyse von auf Wellenformen basierenden Audiodaten ist im Hinblick auf effizientes und effektives Musikretrieval von fundamentaler Bedeutung. Exemplarisch seien an dieser Stelle die Themenkomplexe des Audiomatching und der Strukturanalyse skizziert. Ziel der *Audioidentifikation* ist die Erkennung einer in einer Datenbank enthaltenen Aufnahme anhand eines kurzen Audiofragments. Die Fragestellung des *Audiomatching* kann als Verallgemeinerung der Audioidentifikation angesehen werden. Hierbei besteht die Anfrage aus einem kurzen Audioausschnitt. Ziel ist dann die automatische Identifikation und Extraktion aller zu dieser Anfrage musikalisch ähnlichen Abschnitte, z.B. unabhängig vom Interpreten oder von der Instrumentation, in der gegebenen Datenbank [Abbildung 3].

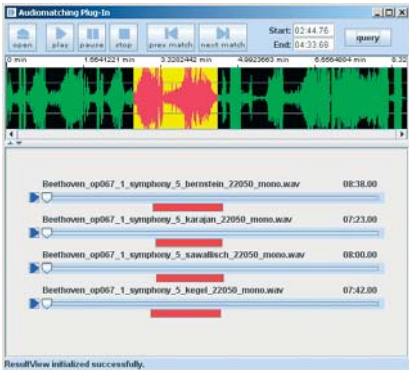


Abbildung 3:
***Oben:* Die Anfrage besteht aus einem Audioabschnitt *(gelber Hintergrund)*, welcher innerhalb einer Wellenformdarstellung ausgewählt werden kann.**
***Unten:* Alle Audioaufnahmen, die Treffer zu dieser Anfrage enthalten, werden inklusive aller Trefferstellen aufgelistet.**

Eine verwandte Fragestellung stellt die *Strukturanalyse* dar, bei der automatisch sich wiederholende Strukturen innerhalb eines Musikstücks (unter Zulassung gewisser musikalischer Variationen) erkannt werden sollen. :::

KONTAKT

Meinard Müller
ABT. 4 Computergrafik
Telefon +49 681 9325-405
Email meinard@mpi-inf.mpg.de

SAPIR: Suche in audio-visuellem Inhalt durch Peer-to-Peer Information Retrieval

SAPIR ist eine experimentelle, über viele Rechner verteilte Suchmaschine, mit der Bilder, Videos und andere audiovisuelle Inhalte – zusammen mit ihren textuellen Beschreibungen und Metadaten wie z.B. GPS-Positionen – indiziert und gefunden werden können.

Die Internetsuche ist heute von wenigen, großen Unternehmen wie Google und Microsoft dominiert; diese verwenden zentralisierte Server-Farmen für die Indexierung und Suche und sind im Wesentlichen auf Suche nach Schlüsselwörtern beschränkt. Selbst die Suche nach audiovisuellen Inhalten ist auf die zu einem Bild oder Video gehörenden Texte und Metadaten beschränkt. Um ein Foto zu finden, müssen die Stichwörter der Anfrage auf derselben Webseite vorhanden sein wie das Foto. So können zum Beispiel der Dateiname des Bildes, der das Bild umgebende Text in der Webseite oder spezielle Wörter in Ankerelementen von Links zu Treffern führen. Eine Suche nach Fotos des Berges Mont Blanc führt daher auch zu Treffern, die Federhalter der gleichnamigen Marke zeigen. Die eigentlichen Bildinhalte, also Farben und Formen im Bild, und die im Bild gezeigten Objekte wie Landschaften, Denkmäler, Personen und anderes werden nicht genutzt.

Ein neueres Suchprinzip für audiovisuelle Daten, das von großen Suchmaschinen bisher nur in sehr beschränktem Kontext verfolgt wird, ist die *Suche durch Beispiel („Query by Example")*. Die Eingabe für eine Suchanfrage nach einem Foto ist einfach selbst ein Foto, und das Result sind inhaltlich ähnliche Bilder. Die Eingabe bei dieser Ähnlichkeitssuche kann man beispielsweise durch eine initiale Schlüsselwortanfrage und Anklicken eines guten Resultats liefern. Auf diese Weise können wir leicht die Treffer zur Suche nach dem Mont Blanc auf Berg- und Gipfelbilder einschränken. Durch die Kombination mit nach wie vor zusätzlich möglichen Schlüsselwörtern oder das Hinzunehmen von GPS-Koordinaten, in deren Umkreis wir suchen wollen, lassen sich Bilder und andere audiovisuelle Inhalte relativ mühelos finden. Beim Wandern, Sightseeing oder auf der Straße kann die initiale Anfrage auch durch das Aufnahmen eines Digitalfotos mit einem Mobiltelefon erzeugt werden oder durch das Summen eines Liedes, für das dann automatisch ähnliche Songs vorgeschlagen werden.

Die SAPIR-Architektur ist speziell für die Anforderungen der effizienten Indexierung von audiovisuellen Inhalten und die Suche durch Beispiel, in Kombination mit Suchbedingungen auf Texten oder Metadaten, entworfen worden. Ein Zusammenschluss vieler Rechner nach dem Peer-to-Peer-Prinzip *(P2P)* zu einem so gennanten P2P-System ermöglicht große Rechenleistung. SAPIR verwendet mehrere solcher P2P-System, jeweils eines für jeweils einen Datentyp: Text, Farbverteilungen in Bildern, Konturen in Bildern, Sprache in Videos, usw.

Ein weiteres P2P-System ist für die Interaktion mit Benutzern und die Zerlegung von Suchanfragen in Teilaufträge für die einzelnen P2P-Systeme zuständig sowie für das Zusammenfügen von Teilergebnissen. In jedem datentypspezifischen P2P-System wird wiederum besonderes Augenmerk auf die Verteilung der Daten auf einzelne Peers und die Verteilung von Anfragen *(„Query Routing")* zu den Peers gelegt. Somit müssen Suchanfragen nur an wenige Peers mit den besten Suchergebnissen weitergeleitet werden. Das Auswählen geeigneter Peers wird anhand raffinierter Statistiken berechnet, die während der Datenindexierung erstellt werden.

Die Praxistauglichkeit, Skalierbarkeit und Effizienz der SAPIR-Architektur wurde in einem Demonstrator realisiert, der u.a. auf der Cebit-Messe 2009 vorgeführt wurde. Dazu wurden als Testfall 100 Millionen Fotos von flickr.com analysiert und indexiert. Suchanfragen können sowohl über einen Web-Browser als auch über eine spezielle Anwendung für Mobiltelefone eingegeben werden. Diese Anwendung ist über die SAPIR-Homepage www.sapir.eu öffentlich zugänglich.

SAPIR wurde im Rahmen eines Forschungsprojekts der Europäischen Union entwickelt. Projektpartner sind zusätzlich zum Max-Planck Institut für Informatik:
IBM Research in Haifa (Israel), die Forschungsorganisation CNR in Pisa (Italien), die Masaryk-Universität in Brno (Tschechien), die Universität in Padova (Italien) und die Industriepartner Eurix (Italien), Xerox (Frankreich), Telefonica (Spanien) und Telenor (Norwegen). :::

KONTAKT

Edwin Lewis-Kelham
ABT. 5 Datenbanken und Informationssysteme
Telefon +49 681 9325-523
Email edwin@mpi-inf.mpg.de
Internet http://www.sapir.eu

TopX 2.0 – Effiziente Suche in digitalen Bibliotheken

Volltext-Suche in semistrukturierten Dokumenten

TopX ist eine Suchmaschine zur Volltextsuche in sogenannten semistrukturierten Dokumenten, einem Datenformat, das zum Beispiel in digitalen Bibliotheken häufig zum Einsatz kommt. Diese Dokumente sind semistrukturiert, da sie semantische Annotationen und Strukturierungsmarken mit reichhaltigen Textpassagen in einem einheitlichen Format, dem bekannten XML-Format (*Extensible Markup Language*), repräsentieren. Insbesondere semantische Annotationen sind interessant für die Anfrageauswertung, da sie entscheidend dazu beitragen können, den textuellen Inhalt, der mit einer solchen Annotation gekennzeichnet wurde, zu disambiguieren und somit besser zu verstehen. So kann zum Beispiel die Anfrage *„Wer waren ehemalige Doktoranden von Max Planck, die selbst bekannte Wissenschaftler wurden?"* in eine sehr spezifische Pfadanfrage über einer entsprechenden Dokumentstruktur übersetzt werden, die dann von TopX effizient und effektiv bearbeitet werden kann:

//article[.//person ftcontains „Max Planck"]
//doctoral_students//scientist
(gemäß XPath 2.0 Full-Text Standard)

Als Treffer dieser Anfrage über einem auf Wikipedia basierenden, annotierten Korpus können in diesem Fall direkte Verweise u.a. zu Gustav Ludwig Hertz, Erich Kretschmann, Walther Meißner geliefert werden, die dann als kompakte Antworten mit weiterführenden Links, zum Beispiel zu deren Enzyklopädieeinträgen, angezeigt werden können.

Explorative Suche

Diese Art der Suche ermöglicht eine sehr viel genauere Formulierung von Anfragen als sie mit einfacher Stichwortsuche – der Funktionalität von Internet-Suchmaschinen – möglich ist. Die Voraussetzung für eine erfolgreiche Anwendung dieser Anfragesprache ist allerdings eine genauere Kenntnis der Dokumentstruktur, des sogenannten XML-Schemas, durch den Benutzer.

Um den Benutzer in der Formulierung solcher Anfragen zu unterstützen, ermöglicht TopX auch eine explorative Form der Suche in digitalen Bibliotheken. Hier kann der Benutzer über einfache Stichwortanfragen eine Dokumentsammlung mit anfangs unbekanntem Schema erkunden, mehr über die enthaltenen Dokumente und deren Schema lernen und dabei schrittweise die Anfragen verfeinern. Die Expressivität der Anfragesprache (die die internationalen W3C-Standards XPath 2.0 und XQuery 1.0 Full-Text unterstützt) erlaubt dabei eine schrittweise Präzisierung der Anfragen. So kann der Benutzer zum Beispiel in einem auf Wikipedia-Artikeln basierenden Korpus mit einer reinen Stichwortsuche nach *„Max Planck"* beginnen. Diese Anfrage würde (unter anderem) Links zu der Person Max Planck als auch zur Max Planck Gesellschaft unter den besten Treffern liefern. Eine einfache Verfeinerung der Anfrage unter Ausnutzung der Dokumentstruktur könnte dann folgendermaßen aussehen

//article[.//person ftcontains „Max Planck"]

und würde der Suchmaschine mitteilen, Treffer zur Person Max Planck zu bevorzugen.

Weitere Möglichkeiten, den Benutzer zu den bestmöglichen Treffern in der Informationssuche zu lenken, bestehen in der dynamischen Relaxierung von Anfragebedingungen sowie in der gezielten Expansion von Anfragen. So können überspezifizierte Anfragen, die sonst zu keinem Treffer führen würden, dynamisch während der Anfrageauswertung relaxiert werden, indem die Anfrage in einen disjunktiven statt konjunktiven Auswertungsmodus wechselt. In einem solchen Fall kann beispielsweise eine sehr spezifische Anfrage nach Bildern, die mit *„Max Planck"* und *„Albert Einstein"* beschriftet sind, auch Treffer liefern, die nur die Stichwörter *„Planck"* und *„Einstein"* enthalten und somit mit hoher Wahrscheinlichkeit ebenfalls relevante Ergebnisse zu dieser Anfrage sind. Darüberhinaus können durch Ontologie-basierte Anfrageexpansionen auch Treffer zu semantisch ähnlichen Konzepten automatisch gefunden werden – zum Beispiel zu *„Albert Einstein"* **an Stelle von** *„Max Planck"*, wenn wir nach deutschen Physikern suchen. Dabei sorgen probabilistische Relevanzmodelle automatisch dafür, dass die besten Anfrageergebnisse direkt als erste Treffer innerhalb einer Rangliste geliefert werden.

Effizienz und Skalierbarkeit

Während der letzten beiden Jahre lag unser besonderer Augenmerk auf einer weiteren Verbesserung der Effizienz in der Anfrageauswertung und der Skalierbarkeit für sehr große Datenmengen. Weite Teile der Implementierung wurden vollkommen neu gestaltet, und die interne Dokumentspeicherung und -indexierung von TopX wurde von der Verwendung eines Datenbanksystems auf eine völlig neu entwickelte, speziell für TopX optimierte Indexstruktur umgestellt. Diese neue Indexstruktur unterstützt hocheffiziente Formen der Indexkomprimierung und des verteilten Indexierens über mehrere Rechner (z.B. eines Clusters) hinweg, was maßgeblich für die Skalierbarkeit der Suchmaschine ist. Unsere experimentellen Ergebnisse mit TopX und die regelmäßige, sehr erfolgreiche Teilnahme an internationalen Benchmark-Wettbewerben untermauern die weltweit starke Position von TopX zur Informationssuche auf semistrukturierten Daten, die kommerzielle Produkte weit übertrifft. :::

KONTAKT

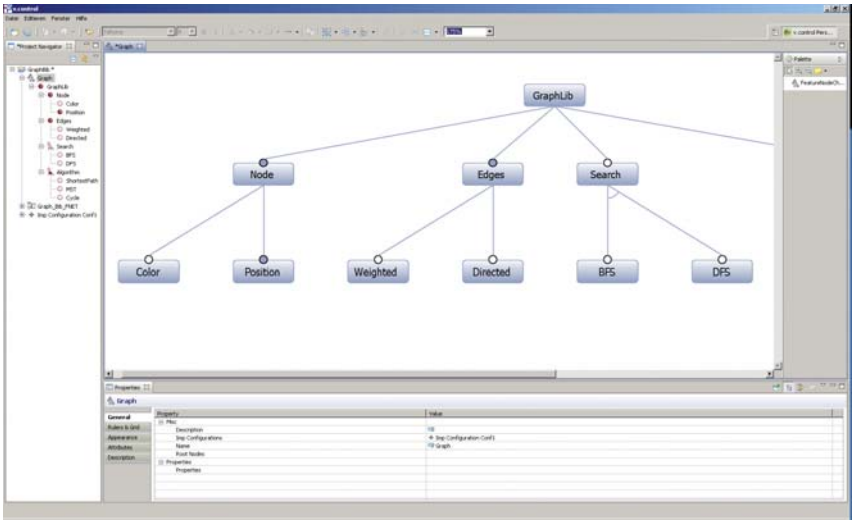
Martin Theobald
ABT. 5 Datenbanken und Informationssysteme
Telefon +49 681 9325-507
Email martin.theobald@mpi-inf.mpg.de

Feature Diagramme

Viele Automobilfirmen bieten heute Konfiguratoren für ihre Produkte im Internet an. Typischerweise stellt ein Kunde ausgehend von der Wahl eines Modells sein persönliches Fahrzeug durch Auswahl möglicher Optionen zusammen. Manche Optionen sind verpflichtend zur Auswahl, wie etwa die Wahl einer Farbe des Autos oder die des Motors, andere sind optional, wie etwa eine Anhängerzugvorrichtung oder der Einbau eines Telefonmoduls. Typischerweise gibt es Abhängigkeiten zwischen den möglichen Auswahlen: ein starker Motor bedingt durch die mögliche Erzielung hoher Geschwindigkeiten bestimmte Radkombinationen, während die Anhängerzugvorrichtung schwache Motoren ausschließt. Feature Diagramme sind ein Formalismus zur Darstellung und Untersuchung solcher Kombinationsmöglichkeiten und Abhängigkeiten.

Die Herstellersicht des Feature Diagramms eines Autos ist sehr viel komplexer, weil sie nicht nur die für den Kunden relevanten Anteile, sondern letztendlich das ganze Auto bis zur einzelnen Schraube beinhaltet. Durch den immer größeren und immer wichtiger werdenden Anteil von Software in einem Fahrzeug ergeben sich zusätzliche Arten von Abhängigkeiten. Softwarekomponenten lassen sich bei weitem nicht so einfach kombinieren wie z.B. Maschinenbauteile, weil Software sich nicht robust verhält.

Weiterhin arbeiten Hersteller heute nicht mehr auf der Grundlage einzelner Modelle, sondern so genannter Produktlinien, bei denen auf der Basis eines möglichst hohen Anteils gemeinsamer Teile verschiedene Produkte realisiert werden. So gehören z.B. der Audi A3 und der VW Golf zur gleichen Produktlinie des VW Konzerns.



Wir untersuchen zusammen mit der PROSTEP IMP GmbH Verfahren, um große Feature Diagramme zu repräsentieren und deren Eigenschaften zu berechnen. Für das Automobilbeispiel wären Beispiele für solche Eigenschaften die Anzahl der möglichen Produkte aus einer Produktlinie unter Vorgabe bestimmter zu verwendender Teilegruppen, oder die Konsistenz von Untergruppen, d.h., ob diese überhaupt noch zu einem Produkt beitragen. Alle diese Probleme sind *„hart“*, d.h. es gibt bis heute kein Verfahren, das sie für beliebig große Feature Diagramme löst. Für das Automobilbeispiel können wir im Moment Diagramme bis zu einer Größe von etwa 6000 Knoten in vernünftiger Zeit behandeln. Das entspricht einem Diagramm, das potentiell 2^{6000} verschiedene Konfigurationsmöglichkeiten repräsentiert, deutlich mehr als die geschätzte Anzahl 2^{1300} der Atome im kompletten Universum. Solch große Probleme lassen sich nur durch die automatische, intelligente Ausnutzung der Abhängigkeitstrukturen in den Diagrammen bearbeiten. Eine komplette Produktlinie entspricht einem Diagramm mit etwa 50000 Knoten. Verfahren für so große Probleme sind Gegenstand der aktuellen Forschung. :::

KONTAKT

Christian Dressler
IT-Abteilung
Telefon +49 681 9325-5847
Email dressler@mpi-inf.mpg.de

Christoph Weidenbach
FG. 1 Automatisierung der Logik
Telefon +49 681 9325-900
Email weidenbach@mpi-inf.mpg.de

VISUALISIERUNG

Der Sehsinn ist einer der wichtigsten Sinne des Menschen. Er ermöglicht es dem Menschen, seine Umgebung sehr schnell zu erfassen. Für den Menschen ist der Sehsinn aber nicht nur ein reines Werkzeug zur Umgebungserfassung. Visuelle Medien wie Gemälde, Fotos, 2D- und 3D- Videos aber auch Filme und Computerspiele haben nicht zuletzt deshalb einen so hohen Stellenwert für uns, weil sie die Vorstellungskraft und das Ästhetikempfinden des Menschen über seinen stärksten Wahrnehmungskanal ansprechen. Im Forschungsbereich Visualisierung entwickeln wir daher Verfahren, mit denen komplexe Informationszusammenhänge analysiert und dargestellt werden können, sowie Verfahren zur schnellen fotorealistischen Darstellung computergenerierter Szenen.

Ein Kernbereich der Visualisierung ist die Analyse und visuelle Darstellung komplexer wissenschaftlicher Datensätze wie sie zum Beispiel bei der medizinischen 3D-Bilderfassung anfallen. Am Max-Planck-Institut für Informatik werden neue Verfahren entwickelt um feinste Strukturen, wie zum Beispiel der Verlauf von Nervenfasern, in solchen 3D-Bilddatensätzen automatisch zu erkennen und sie entsprechend darzustellen. Mit Hilfe solcher Verfahren kann der Arzt schneller und zuverlässiger eine Diagnose stellen.

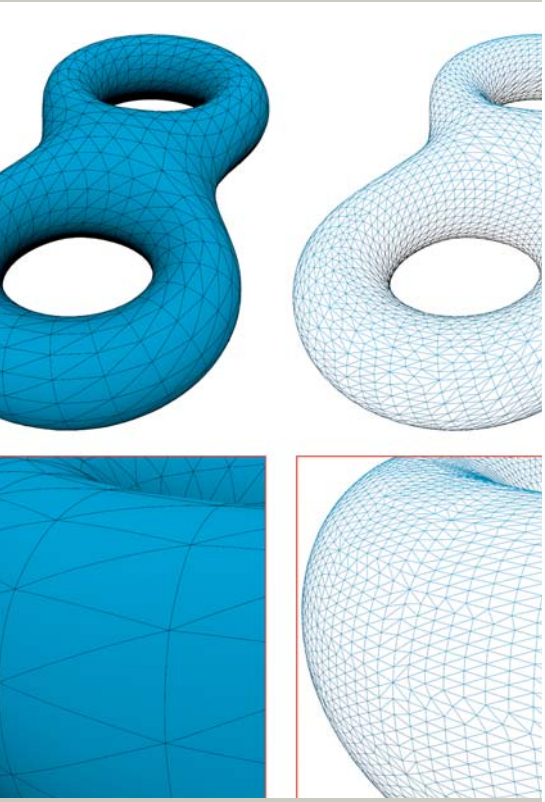
Um virtuelle Welten naturgetreu erscheinen zu lassen, wird am Max-Planck-Institut für Informatik an Methoden zur akkuraten Echtzeitsimulation der Lichtausbreitung in Szenen geforscht. Die schnelle Durchführung solch komplexer Berechnungen ist eine der großen Herausforderungen der Computergrafik. So wurden zum Beispiel neue Verfahren entwickelt um weiche Schatten sehr schnell darzustellen.

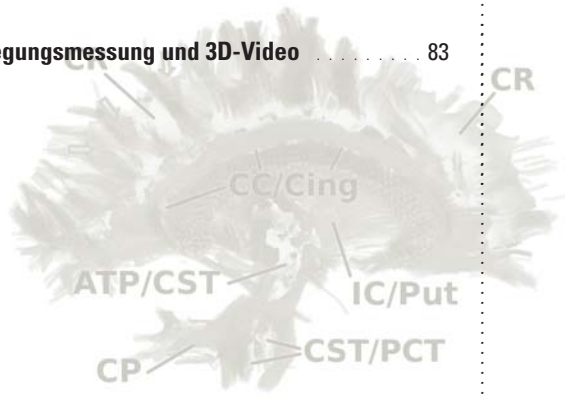
Heutzutage gibt es neben Computern eine Vielzahl an weiteren Geräten, die dreidimensionale Grafiken darstellen können. Die verschiedenen Gerätetypen unterscheiden sich sehr in ihrer Fähigkeit, Bildinhalte darzustellen, da sie unterschiedliche Displaygrößen, unterschiedlich leistungsfähige Prozessoren, oder unterschiedlich große Speicher besitzen. Am Max-Planck-Institut für Informatik werden daher Algorithmen und Datenstrukturen erforscht, mit deren Hilfe die gleichen 3D-Szenen adaptiv auf unterschiedlichen Plattformen dargestellt werden können. In anderen Worten, die Inhalte und die Darstellung werden den Geräten angepasst. In diesem Zusammenhang werden auch neue Algorithmen entwickelt, um extrem große Datensätze auch auf Geräten mit begrenztem Hauptspeicher anzuzeigen.

Neue Bildsyntheseverfahren erstellen Grafiken mit einem sehr hohen dynamischen Bereich (High Dynamic Range – HDR), das heißt die Spannweite der Helligkeitswerte liegt weit über dem was von normalen Displaysystemen dargstellt werden kann. Am Max-Planck-Institut für Informatik werden daher neue Algorithmen zur Bildverarbeitung im HDR Bereich entwickelt, in dem viele Operationen genauer ausgeführt werden können. Zudem wird unter Ausnutzung von Modellen der menschlichen Wahrnehmung an Verfahren gearbeitet, um HDR Bilder auf Displays mit begrenztem dynamischem Bereich darzustellen.

Nicht nur die Darstellung von komplexen 2D- und 3D-Szenen, sondern auch deren Modellierung ist eine Herausforderung. Am Max-Planck-Institut für Informatik wird an Methoden zur datengetriebenen Modellierung gearbeitet. Anstatt Szenenmodelle von Hand zu konstruieren werden hierbei Modelle aus Bild- und Videodaten der realen Welt gelernt. Eines dieser neuen Verfahren ermöglicht es, das Fließverhalten von Flüssigkeiten aus Videos zu lernen und auf andere Bildinhalte zu übertragen. In einem weiteren Forschungsschwerpunkt werden neue Algorithmen entwickelt, um realistische dynamische 3D-Modelle von Schauspielern aus Videodaten zu rekonstruieren. :::

BEITRÄGE





Adaptive Bildsynthese auf unterschiedlichen Plattformen

Die Erstellung von visuellen Inhalten ist einer der großen Kostenfaktoren in vielen Branchen, seien es Filme, Computerspiele, Simulationen, Zeitungen oder wissenschaftliche Abhandlungen. Viel künstlerisches Können, technische Versiertheit oder auch Kreativität sind von Nöten, um ein überzeugendes Resultat zu erzielen. Dies ist besonders problematisch in der heutigen Zeit, in der die Ausgabegeräte für Informationen vielfältig sind. Die Berücksichtigung dieses Aspekts erfordert Lösungen, die sich dem Medium entsprechend anpassen. Es ist nahezu unmöglich eine Darstellung eigens für jedes mögliche Ausgabegerät zu erstellen und hier setzt unsere Arbeit an.

Wir untersuchen, wie visuelle Aspekte adaptiv gemacht werden können, um sie auf unterschiedlichen Plattformen in bestmöglicher Qualität darzustellen. Letzteres kann sowohl den Inhalt selber, als auch den Algorithmus betreffen, der zur Darstellung notwendig ist. Die Bandbreite ist weit; von intelligenten Inhalten, wie einem Bild in einer Präsentation, welches seine Farbpalette automatisch der Restpräsentation angleicht, zu einem Algorithmus zur Berechnung von Schatten in einer dreidimensionalen Umgebung, dessen Präzision sich dem Prozessor der Maschine anpasst.

Adaptive Visualisationen

Heutzutage haben wir eine Vielzahl an unterschiedlichen Ausgabemöglichkeiten, die allesamt verschiedene Eigenschaften aufweisen. Handys haben sehr kleine Bildschirme, Beamer liefern riesige Projektionen. HDR erlaubt hohen Kontrast, während diese Möglichkeit auf normalen Bildschirmen nicht vorhanden ist.

Eines unserer Projekte, welches diese Forschungsrichtung illustriert, ist die Darstellung von Vektorgrafik. Diese ist auflösungsunabhängig, welches sie problemlos skalieren lässt. Desweiteren haben wir Lösungen, die eine Stilisation ermöglichen. Ein Inhalt kann somit sehr variable Erscheinungsbilder haben. Die Illustration [Abbildung 1] zeigt verschiedene Vektorbilder, die automatisch stilistisch angepasst wurden. Die Auswahl des Stils könnte von einem Gerät oder Dokument direkt bestimmt werden.

Abbildung 1: Automatisch generierte Vektorillustrationen in verschiedenen Stilen

Skalierbare Algorithmen

Realistische Bildsynthese von virtuellen Welten ist ein kompliziertes Unterfangen von hohem Stellenwert. Zu Simulationszwecken ist es wichtig, dass die erstellten Bilder realistisch erscheinen. Hierzu sind komplexe Berechnungen nötig, die physikalische Elemente, wie Beleuchtung (Schatten, Spekularitäten, etc.), Materialien (Semitransparenz, Reflektanz, ...) oder Aufnahmeapparate (Tiefenunschärfe, Linsenreflexionen, etc.) simulieren. Abbildung 2 zeigt einige komplizierte Effekte, die nur mit Hilfe von Approximationen in Echtzeit dargestellt werden können. Um so wichtiger ist es in der Lage zu sein, die aus den Annäherungen resultierenden Fehler zu verstehen und zu begrenzen. Nur unter solchen Voraussetzungen ist es möglich ein bestmögliches Bild zu erstellen, ohne dabei ein bestimmtes Budget zu überschreiten.

Abbildung 2: Sanfte Schatten, Tiefenunschärfe oder Transluzenz sind kostspielige Effekte, die nur durch optimierte Algorithmen echtzeitfähig werden.

Das Problem der Darstellung beginnt bereits bei der Datenmenge. Heutzutage arbeiten wir mit riesigen Datensätzen von hunderten von Gigabyte. Diese entstehen z.B. durch 3D-Scanner, medizinische (Computer- und Magnetresonanztomografie) oder prozedurale Verfahren. Nur durch neue Repräsentationsformen und Datenstrukturen wird es möglich eine Darstellung auf dem Computer zu ermöglichen.

Abbildung 3 zeigt Echtzeitvisualisierungen von mehreren hundert Gigabyte an Daten. Solch eine skalierfähige Bildsynthese könnte die Zukunft der Computergrafik maßgeblich mitbestimmen. :::

Abbildung 3: Unsere Algorithmen erlauben eine interaktive Visualisierung von Gigabytes an Daten und machen Semitransparenz und große Detailfülle möglich.

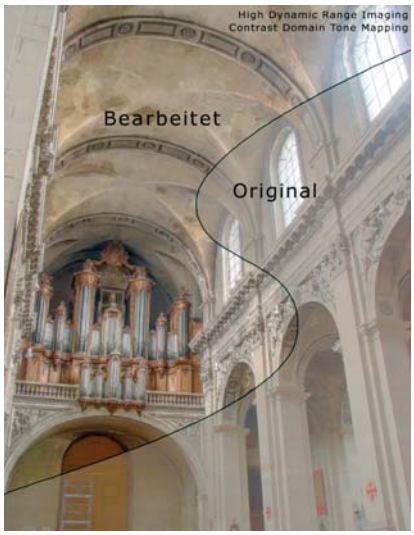
KONTAKT

Elmar Eisemann
ABT. 4 Computergrafik
Telefon +49 681 9325-416
Email eisemann@mpi-inf.mpg.de

HDR – Bilder und Videos mit erhöhtem Kontrastumfang

Die meisten Bild- und Videokameras können nur einen gewissen Teil des Farb- und Helligkeitsspektrums *(high dynamic range)* speichern. Das Auge sieht deutlich mehr Farben und Kontraste. Außerdem sind viele dieser Bilder und Filme qualitativ nicht hochwertig genug, um auf den Bildschirmen der neuesten Generation dargestellt zu werden. So wurde zum Beispiel das weit verbreitete Bildformat JPEG aus Effizienz-Gründen entwickelt: Im JPEG werden nur so viele Informationen gespeichert, wie auf Standardbildschirmen und Ausgabegeräten, die zum Entwicklungszeitpunkt des JPEG-Formats existierten (Kathodenstrahlrohr-Bildschirme und Fernsehgeräte), wiedergegeben werden können. Diese Annahme ist aber nicht mehr aktuell. Heutige LCD- und Plasmabildschirme können eine viel größere Farbskala *(color gamut)* und einen größeren Helligkeitsbereich *(dynamic range)* darstellen als ihre Vorgänger.

Das Tone Mapping komprimiert die hohe Dynamik *(rechter Einsatz)*, indem es das Kontrastverhältnis zwischen den Zonen gleicher Belichtung im zerlegten HDR Bild *(linker Einsatz)* optimiert. Das Ergebnis zeichnet sich durch gute Tonabbildung in allen Bildbereichen aus.



Rendern von HDR-Bildern für Geräte mit limitiertem Kontrastumfang

Die „High Dynamic Range" *(HDR)* Bildverarbeitung *(HDRI)* überwindet die Grenzen der bisherigen Bildverarbeitung, indem alle Farboperationen sehr viel genauer ausgeführt werden. Selbst die Wahrnehmung des menschlichen Auges wird übertroffen. Mit HDRI lassen sich Bilder von natürlichen Szenen mit allen wahrnehmbaren Farben ohne Unter- oder Überbelichtung originalgetreu wiedergeben. HDRI arbeitet also nicht nur präziser, sondern schafft es auch, visuelle Signale in der menschlichen Perzeption zu synthetisieren oder zu visualisieren. So können im Gegensatz zu traditionellen Bildern die HDR-Bilder visuelle Phänomene wie selbstleuchtende Oberflächen (Sonne, Glühlampen), Glanzpunkte, Schatten sowie lebendige, stark gesättigte Farben darstellen.

In der Forschungsgruppe der Computergrafik haben wir neue Video- und Bildformate entwickelt, die natürliche Szenen äußerst genau kodieren. Um die größeren Datenmengen entsprechend verwalten zu können, erreichen diese Formate gute Kompressionsraten. Wir haben außerdem ein Softwarepaket entwickelt *pfstools* http://pfstools.sourceforge.net/, das nützliche Programme für die HDR-Bildverarbeitung beinhaltet und für zukünftige Forschungsarbeiten gedacht ist. Mit unserer Arbeit haben wird versucht, das Softwarepaket möglichst unabhängig von spezieller Bildtechnologie zu gestalten. Auf diese Weise schränken uns nur die Fähigkeiten der menschlichen Perzeption ein. Bereits bestehende Bildverarbeitungssoftware und Hardware auf HDR-Daten umzugestalten, erfordert jedoch viel Aufwand. Auch waren einige neue Definitionen von Standards für die Bildverarbeitung notwendig. Im Wesentlichen wollen wir das HDR-Bildverarbeitungs-Konzept popularisieren, neue Standardwerkzeuge und Algorithmen für die Verarbeitung von HDR-Daten entwickeln und die Forschung im Bereich visueller Perzeption vorantreiben, da diese einen entscheidenden Einfluss auf die digitale Bildverarbeitung hat. :::

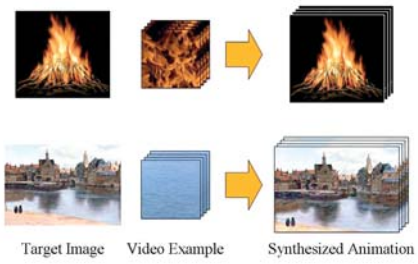
KONTAKT

Karol Myszkowski
ABT. 4 Computergrafik
Telefon +49 681 9325-429
Email karol@mpi-inf.mpg.de
Internet http://www.mpi-inf.mpg.de/resources/hdr/

Animierte Darstellung von Flüssigkeiten unter Verwendung von Videobeispielen

Bei der Betrachtung eines eine natürliche Szene darstellenden Photos oder Bildes stellen wir uns häufig vor, wie diese in der Realität aussehen könnte. Bewegungen spielen dabei eine bedeutsame Rolle. Es ist jedoch ein ungelöstes Problem, wie Flüssigkeiten und Gase in unbewegten Bildern überzeugend animiert werden können. Die Möglichkeit dieser Bildmanipulation würde neue, wichtige Anwendungsfelder erschließen. So werden zum Beispiel in der digitalen Filmbranche oft bewegte Szenen als Hintergrund verwendet um die dargestellte Szene realistisch und „in Bewegung" erscheinen zu lassen.

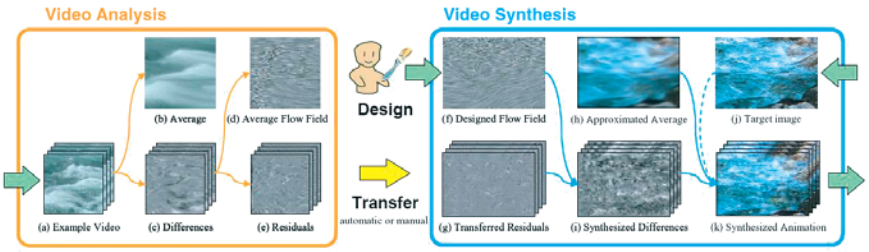
Wir haben ein System entwickelt, das es einem Benutzer ermöglicht, ein statisches Bild – welches Strömungselemente enthält – realistisch zu animieren. Unser Algorithmus benötigt zwei Eingabeparameter: 1. ein Beispielvideo, das die Charakteristik der zu erzeugenden Strömung gut wiedergibt, nicht notwendigerweise jedoch die richtigen Farben, Texturen oder auch Strömungsrichtungen aufweist, und 2. ein vom Benutzer grob skizziertes Strömungsfeld, das die Richtung der zu erzeugenden Animation vorgibt. Das manuell spezifizierte Strömungsfeld wird automatisch mit Hilfe eines Gradientenverfahrens an das statische Zielbild angepasst. Nach diesem Vorverarbeitungsschritt kann der Benutzer hochfrequente Videoinformation halbautomatisch vom charakteristischen Video (1.) auf das statische Zielbild übertragen. Mit Hilfe unseres Algorithmus und einer Videodatenbank können überzeugende Animationen von Flüssigkeiten und Gasen in kurzer Zeit realisiert werden ohne auf teure Simulationen zurückgreifen zu müssen.



Unser Algorithmus ermöglicht Anwendern die Generierung kontinuierlicher Bewegungsabläufe von Flüssigkeiten und Gasen. Dazu werden nur ein statisches Bild der zu animierenden Szene und ein Beispielvideo einer beliebigen anderen Strömung benötigt. Das System erzeugt dann halb-automatisch überzeugende, animierte Versionen des statischen Originalbildes.

Analyse und Darstellung von animierten Flüssigkeiten

Unser System arbeitet wie folgt: Ein Videobeispiel *(a)* wird in drei Komponenten zerlegt. Diese sind ein Mittelwertbild *(b)*, ein Strömungsvektorfeld *(d)* und verbleibende Fehlerterme (Residuen) *(e)*. Das Strömungsfeld wird mit Hilfe des optischen Flusses bestimmt. Jede Abbildung einer Strömung kann in diese drei Teile zerlegt werden, d.h. dass eine Strömungsanimation dadurch erzeugt werden kann, dass geeignete drei Komponenten für das Zielbild definiert werden. Da es ein schweres Problem ist diese Definition vollautomatisch vorzunehmen, enthält unser System eine Benutzerumgebung mit deren Hilfe es auf einfache Art und Weise möglich ist das Strömungsfeld *(f)* zu erzeugen, Residuumskomponenten des Videobeispiels auf das Zielbild zu transferieren *(g)*, und das Mittelwertbild zu approximieren *(h)*. Unser Algorithmus ist mit diesen Eingaben in der Lage realistische Animationen des Zielbildes *(k)* zu erzeugen. :::



Systemübersicht, grüne Pfeile repräsentieren die Eingabe und Ausgabe des Systems.

KONTAKT

Makoto Okabe
ABT. 4 Computergrafik
Telefon +49 681 9325-427
Email mokabe@mpi-inf.mpg.de

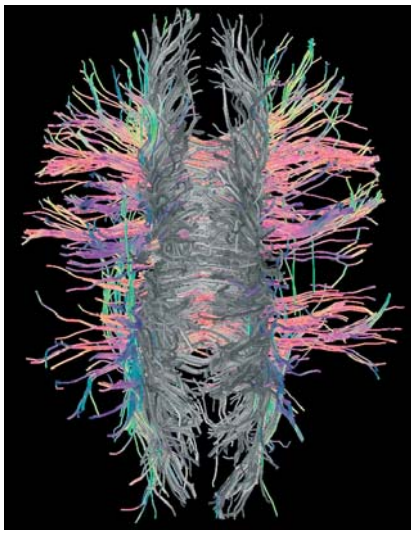
Merkmalsbasierte Visualisierung von Daten der Diffusions-Bildgebung

Mit der Diffusions-Bildgebung steht ein bildgebendes Verfahren zur Verfügung, das eine nichtinvasive Untersuchung der Nervenfaser-Verbindungen im menschlichen Gehirn ermöglicht. Eine Variante der Magnetresonanztomographie *(MRT)* misst hierbei die natürliche Wärmebewegung von Wassermolekülen im Gehirn. Da diese quer zur Richtung großer Nervenbündel stark eingeschränkt ist, lässt sie Rückschlüsse auf Ort und Ausrichtung der Nervenverbindungen zu. Anwendung finden die so gewonnen Daten unter anderem zur Erforschung neurologischer Erkrankungen und normaler Hirnfunktionen sowie bei der Planung neurochirurgischer Eingriffe.

Um die richtungsabhängige Beweglichkeit der Moleküle zu erfassen, ist in jedem Volumenelement *(Voxel)* ein Diffusionsmodell erforderlich, das durch zahlreiche Parameter beschrieben wird. Dies führt zu großen Datenmengen, deren vollständige visuelle Darstellung das menschliche Fassungsvermögen überfordern würde. Das Ziel unseres Forschungsprojekts war es daher, in einer automatischen Vorverarbeitung aus den umfangreichen Daten Merkmale zu extrahieren, die anatomisch relevante Strukturen erfassen und eine Abstraktion der Messdaten bilden, die sinnvoll visualisiert werden kann.

Rekonstruktion von Faserbündeln

Traktographie-Methoden rekonstruieren aus Daten der Diffusions-Bildgebung den Verlauf von Nervenbahnen zwischen verschiedenen Hirnzentren. Eine Einschränkung dieser Verfahren liegt in der beschränkten Auflösung der Daten: Die Voxel sind typischerweise deutlich größer als einzelne Nervenzellen, so dass die Messwerte prinzipiell einen Mittelwert über ganze Faserbündel darstellen. Die Interpretation der resultierenden Daten ist vergleichsweise einfach, solange in einem Voxel eine einzige Faserrichtung vorherrscht. Berühren oder kreuzen sich dagegen mehrere Bündel, ist es zur Interpretation des gemeinsamen Signals erforderlich, ihre jeweiligen Anteile zu isolieren.



Die Rekonstruktion der farbig dargestellten lateralen Faserbündel erforderte unsere Tensor-Approximation.

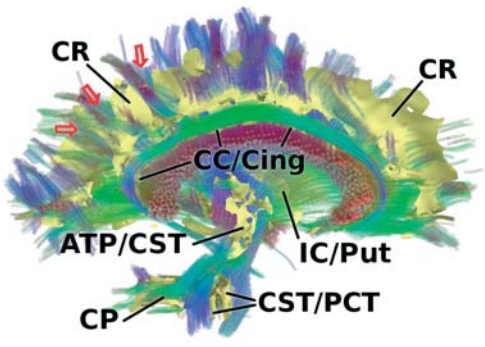
Ein Beitrag unserer Forschung besteht in einer Methode, um die Traktographie auch in solch komplexen Voxeln zuverlässig fortsetzen zu können. Sie nutzt Konzepte der multilinearen Algebra, indem sie die Orientierungsdichtefunktionen, die häufig zur Beschreibung des Diffusionsverhaltens eingesetzt werden, als Tensoren höherer Stufe auffasst. Diese approximiert sie anschließend durch Tensoren niedrigen Rangs, wobei jeder Rang-1-Term dem Anteil eines Faserbündels entspricht. Hierdurch gelingt es, systematische Fehler zu vermeiden, die in zahlreichen früheren Arbeiten auftraten und die Rekonstruktion bestimmter Faserbündel verhinderten.

Ein weiteres Augenmerk lag auf der gemeinsamen visuellen Darstellung von Traktographie-Ergebnissen und herkömmlichen MRT-Daten. In Kooperation mit dem Max-Planck-Institut für Kognitions- und Neurowissenschaften in Leipzig ahmten wir algorithmisch die Klingler-Präparationsmethode nach und konnten auf diese Weise Darstellungen erzeugen, die räumliche Verhältnisse zwischen beiden Datensätzen verdeutlichen.

Topologische Merkmale

Ein zweiter Teil unserer Forschung orientierte sich an topologischen Merkmalen in Vektorfeldern, die sich für die visuelle Analyse komplexer Strömungsfelder als geeignet erwiesen haben. Wir stellten fest, dass eine unmittelbare Übertragung der mathematischen Definitionen auf Diffusions-Tensor-Daten leider nicht zu aussagekräftigen und stabilen Merkmalen führt. Dennoch gelang es uns, aus der anatomischen Bedeutung der Daten heraus neuartige topologische Merkmale zu entwickeln, die auch der Unsicherheit der aus Diffusions-Daten gewonnenen Verbindungsmaße Rechnung tragen.

Anisotropie-Maße quantifizieren, inwiefern die beobachtete Beweglichkeit der Moleküle richtungsabhängig ist. Extremalflächen dieser Maße wurden unlängst als eine Möglichkeit vorgestellt, strukturelle Skelette aus Diffusions-Daten zu extrahieren. Auch zu diesem Forschungszweig konnten wir einen Beitrag leisten, indem wir die topologischen Eigenschaften der Flächen aufgeklärt und einen verlässlichen und effizienten Algorithmus zu ihrer Extraktion entwickelt haben. :::



Eine anatomisch annotierte Anisotropie-Extremalfläche *(gelb)* im Kontext einer Traktographie der linken Hemisphäre

KONTAKT

Thomas Schultz
ABT. 4 Computergrafik
Telefon +49 681 9325-400
Email schultz@mpi-inf.mpg.de

Steigerung des Realismus im Echtzeitrendern

Ein Kernanliegen der Computergrafik ist das Generieren synthetischer Bilder. Hierbei strebt man häufig danach, Ergebnisse zu erzielen, die realistisch wirken und von einem Foto nicht zu unterscheiden sind. Während dies prinzipiell bereits möglich ist, stößt man in der Praxis auf viele Herausforderungen. So ist einerseits die darzustellende Szene mitsamt ihrer Geometrie und den Materialien detailliert zu beschreiben. Andererseits erweist sich die Berechnung des finalen Bildes als komplex und aufwendig, sodass diese schnell viel Zeit in Anspruch nehmen kann.

Dies konfligiert aber insbesondere mit dem Wunsch, Bilder in Sekundenbruchteilen synthetisieren und anzeigen zu können, und somit eine interaktive Erkundung virtueller Szenen zu ermöglichen, was für eine Vielzahl von Anwendungen, wie etwa Computerspiele, von zentraler Bedeutung ist. Im entsprechenden Teilgebiet, dem Echtzeitrendern, bei dem Geschwindigkeit vorrangiges Ziel ist, ist man folglich gezwungen, Abstriche bei der Genauigkeit der Ergebnisse sowie dem realistischen Aussehen zu machen.

Die aktuelle Forschung beschäftigt sich daher mit der Entwicklung neuer Verfahren, die die in Echtzeit erzielbare visuelle Qualität verbessern helfen. Viele Lösungen bedienen sich dabei neuester Grafikhardware, wie sie heutzutage in fast jedem modernen Computer zu finden ist. Diese übertrifft die Rechenleistung aktueller CPUs um ein Vielfaches, erfordert aber auch spezifische Techniken und Formulierungen, um diese nutzen zu können.

Flächenlichtquellen führen zu weichen Schatten, deren Halbschattenbereiche je nach Distanz zwischen Schattenwerfer und -empfänger in ihrer Breite variieren.

Weiche Schatten

Eines der Teilprobleme, die intensiv untersucht werden, ist die Berechnung qualitativ hochwertiger Schatten. Diese tragen entscheidend zum Realismus bei und liefern wichtige Hinweise für die Wahrnehmung.

Dabei liegt der Fokus auf Flächenlichtquellen, wie sie etwa in Räumen anzutreffen sind. Diese liefern weiche Schatten mit weichen Übergängen vom Kernschatten in den schattenlosen Bereich. Dagegen erzeugen die im Echtzeitrendern üblicherweise eingesetzten, idealisierten Punktlichtquellen lediglich harte Schatten, die nicht zuletzt aufgrund der fehlenden Halbschattenbereiche oft unrealistisch wirken.

Um den Grad der Verschattung eines Szenenpunktes zu berechnen, muss letztlich der Anteil der Lichtquelle bestimmt werden, der von diesem Punkt aus sichtbar, also durch Schattenwerfer unverdeckt ist. Während dies bei Punktlichtquellen effizient und schnell möglich ist, da nur die binäre Sichtbarkeit eines Lichtpunktes zu betrachten ist, erfordert die Schattenberechnung für Flächenlichtquellen ausgefeilte Approximationen um möglichst genaue Ergebnisse noch in Echtzeitgeschwindigkeit zu erreichen.

Gekrümmte Flächen

Realismus zeigt sich außer bei Beleuchtungseffekten beispielsweise auch in der verwendeten Szenengeometrie. Viele Objekte des Alltags sind sichtbar glatt, etwa eine Kaffeetasse oder die Karosserie eines Autos. Deren Oberflächen lassen sich mittels gekrümmter Flächenprimitive, wie etwa Bézierdreiecken, beschreiben. Diese erlauben im Gegensatz zu Dreiecksnetzen eine kompakte Repräsentation, erleichtern das Modellieren und Animieren und sind auflösungsunabhängig.

Anderseits sind gekrümmte Flächenstücke nicht direkt darstellbar. Daher werden diese beim Rendern zunächst durch ein Dreiecksnetz approximiert, welches dann angezeigt wird. Diese Umwandlung, Tesselierung genannt, erfolgt in Echtzeit und an die jeweilige Ansicht angepasst, sodass die Dreiecke gerade klein genug sind, um ein visuell glattes Erscheinungsbild zu erzeugen. :::

Oberflächen lassen sich mit gekrümmten Flächenstücken beschreiben *(links)*. Zur Darstellung werden diese durch Dreiecke approximiert *(rechts)*. Diese Umwandlung erfolgt adaptiv, sodass die Oberfläche für jede beliebige Ansicht visuell glatt erscheint.

KONTAKT

Michael Schwarz
ABT. 4 Computergrafik
Telefon +49 681 9325-418
Email mschwarz@mpi-inf.mpg.de

Markerlose optische Bewegungsmessung und 3D-Video

Rekonstruktion detaillierter Animationsmodelle aus Multivideo Daten

Virtuelle Schauspieler, so genannte Avatare, haben sich zu einem wichtigen Bestandteil visueller Medien entwickelt, so zum Beispiel in Filmen, Computeranimationen oder virtuellen vernetzten Welten. Um eine virtuelle Person überzeugend darzustellen, müssen die einzelnen Aspekte der Person, wie zum Beispiel ihre Bewegung, ihre Geometrie und ihre texturelle Erscheinung, extrem realistisch dargestellt werden. Eine Möglichkeit um dies zu erreichen besteht darin, jeden Teilaspekt dieses Gesamterscheinungsbildes manuell in einem Animationsprogramm zu definieren. Dies ist allerdings ein extrem zeitaufwendiger und komplexer Prozess. Die Geometrie der Person muss in präziser Detailarbeit konstruiert werden und jede Facette der Bewegung muss fein abgestimmt werden. Es ist daher kaum erstaunlich, dass komplett manuell erzeugte Animationen, insbesondere hinsichtlich der Qualität der Bewegungsanimation, nicht den Detailgrad eines echten Menschen erreichen.

Die Alternative zur vollständig manuellen Modellierung besteht darin, Teilaspekte der Animation an echten Menschen zu messen. So ermöglichen es Motion Capture Systeme, die Bewegung eines Skelettmodells aus Videobildströmen einer sich bewegenden Person zu rekonstruieren. Leider ist diese Bewegungserfassung ein sehr komplexer Prozess und die zu vermessende Person muss oftmals einen speziellen hautengen Anzug mit optischen Markierungen tragen. Zudem kann mit solchen Systemen die sich über die Zeit verändernde Geometrie oder Textur einer Person nicht gemessen werden.

Wir haben daher einen neuartigen *Performance Capture Algorithmus* entwikkelt, der aus nur acht Multivideoströmen detaillierte Bewegung, dynamische Geometrie und dynamische Textur einer Person in komplexer Kleidung, zum Beispiel einem Rock oder einem Ballkleid, rekonstruieren kann [Abbildung].

Ein Bild aus einer Multivideosequenz, die mit acht Kameras aufgezeichnet wurde. *Rechts:* Das mit unserem Performance Capture Algorithmus rekonstruierte Geometriemodell der Person in der gleichen Pose.

Kernbestandteil der Methode ist ein neuartiges Trackingverfahren für deformierbare Oberflächenmodelle. Unser System erfordert keine optischen Markierungen in der Szene und erfasst dennoch die dynamische Geometrie in hoher Auflösung, wie zum Beispiel die Falten in einem Rock [Abbildung].

Wir haben auch Methoden entwikkelt, welche aus den gemessenen Performancedaten ein Vorwärtssimulationsmodell schätzen. Das bedeutet, wir identifizieren vollautomtisch Kleidungssegmente in der Szene und schätzen deren Materialparameter. Somit können durch einfache Veränderung der Körperbewegungsparameter neue Animationen erstellt werden, in denen sich die Oberflächen hochrealistisch deformieren.

3D-Video und reflektanzbasierte Rekonstruktion detaillierter dynamischer Szenengeometrie

Mit der im vorherigen Abschnitt beschriebenen Methode kann ein detailliertes dynamisches Geometriemodell einer Person gemessen werden. Da es sich bei den Eingabedaten um optisch nicht veränderte Videoströme handelt, kann auch eine dynamische Oberflächentextur der Person geschätzt werden. Hierzu kann man zum Beispiel die Kamerabilder auf das Geometriemodell rückprojizieren *(projective texturing)* und die Pixeldaten auf der Oberfläche interpolieren. Auf diese Weise kann ein so genanntes 3D-Video einer Person gerendert werden, dass heisst man kann in Echtzeit einen beliebigen Kamerablickwinkel auswählen und sieht ein realistisches Abbild der Person aus dieser neuen Perspektive.

Zur Erzeugung von 3D-Videos, die auch unter neuen Beleuchtungsbedingungen dargestellt werden können, haben wir ein Verfahren entwickelt, um ein Reflektanzmodell eines jeden Punktes auf der Körperoberfläche zu schätzen. Hierzu wird die Person mit mehreren kalibrierten Videokameras und unter kalibrierter Beleuchtung aufgenommen. Aus diesen Bilddaten schätzt anschließend ein Optimierungsverfahren parametrische Reflektanzmodelle auf der Oberfläche. Mit Hilfe dieser Modelle kann nun das Erscheinungsbild der Person unter neuer simulierter Beleuchtung realistisch dargestellt werden.

Die Reflektanzbeschreibung kann nicht nur zur Darstellung, sondern auch zur Verbesserung der Geometrieschätzung eingesetzt werden. Ein neuartiges von uns entwickeltes fotometrisches Raumzeit-Stereoverfahren kann selbst feinste Details auf der Oberfläche messen und erreicht damit einen geometrischen Detailgrad, der weit über dem liegt, der mit dem ursprünglich beschriebenen Performance Capture Algorithmus erzielt werden kann. :::

KONTAKT

Christian Theobalt
ABT. 4 Computergrafik
Telefon +49 681 9325-428
Email theobalt@mpi-inf.mpg.de
Internet http://www.mpi-inf.mpg.de/resources/perfcap/

International Max Planck Research School for Computer Science (IMPRS-CS)

Die Ausbildung des wissenschaftlichen Nachwuchses ist von elementarer Bedeutung für die Zukunft von Wissenschaft, Forschung und Innovation in Deutschland. Die Max-Planck-Gesellschaft hat daher gemeinsam mit der Hochschulrektorenkonferenz eine Initiative zur Nachwuchsförderung ins Leben gerufen: die International Max Planck Research Schools (IMPRS). Diese bieten besonders begabten deutschen und ausländischen Studierenden die Möglichkeit, sich im Rahmen einer strukturierten Ausbildung unter exzellenten Forschungsbedingungen auf die Promotion vorzubereiten. Auf diese Weise sollen verstärkt junge Wissenschaftler angeworben und ausgebildet werden.

IMPRS-CS

Förderung des Wissenschaftlichen Nachwuchs

Die IMPRS-CS ist ein Angebot für Nachwuchswissenschaftlerinnen und -wissenschaftler, die zwischen dem Bachelor- oder Masterabschluss und der Promotion stehen. Dies umfasst ein erstklassiges Ausbildungsangebot, wissenschaftliche Schwerpunktbildung, oft mit thematischer Verzahnung mehrerer Promotionen, und eine enge Zusammenarbeit von Doktoranden und ihren Betreuern.

Ein Schwerpunkt liegt auf der internationalen Zusammenarbeit: Die IMPRS-CS will insbesondere ausländische Bewerberinnen und Bewerber für eine Promotion in Deutschland gewinnen, sie mit den Forschungseinrichtungen vertraut machen und ihr Interesse für eine spätere Tätigkeit in oder in Kooperation mit deutschen Forschungseinrichtungen wecken. Über 50 Prozent unserer Doktoranden stammen aus dem Ausland, wobei Bulgarien, China, Indien und Rumänien zu den am stärksten vertretenen Herkunftsländern zählen.

Programme der IMPRS-CS

Die IMPRS-CS bietet in Zusammenarbeit mit der Universität des Saarlandes und der Saarbrücker Graduiertenschule für Informatik Programme für die Qualifizierung zum Promotionsstudium sowie für die Promotion selbst.

Alle Graduiertenprogramme werden in enger Kooperation mit dem Max-Planck-Institut für Informatik, dem Max-Planck-Institut für Softwaresysteme und dem Fachbereich Informatik der Universität des Saarlandes angeboten. Die Projekte werden gemeinsam von den Wissenschaftlern der Max-Planck-Institute und deren Kollegen aus dem Fachbereich Informatik der Universität betreut. Hervorragende Englischkenntnisse sind für alle Bewerber unerlässlich.

Finanzielle Unterstützung

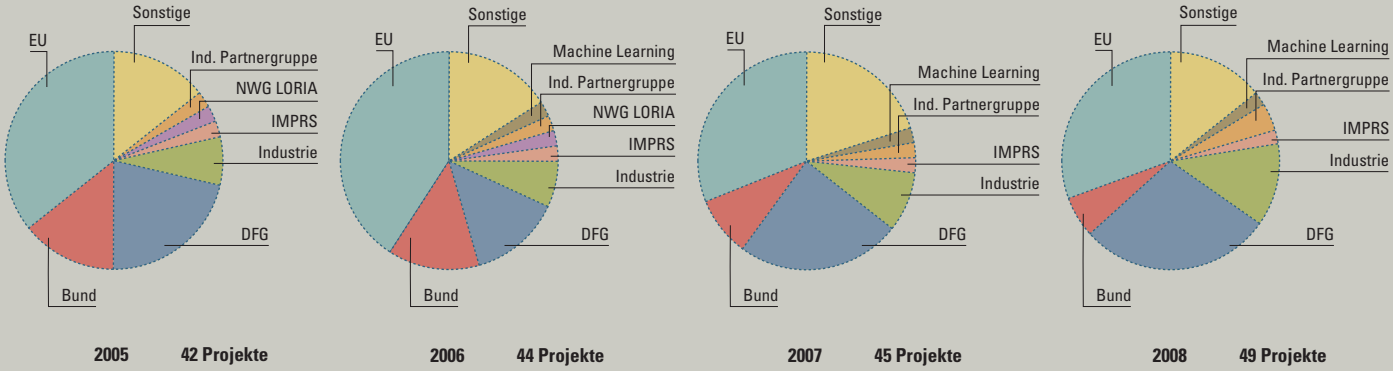
Die zur IMPRS-CS zugelassenen Studierenden erhalten ein Stipendium, das Gebühren, Lebenshaltungskosten und Krankenversicherungskosten sowohl der Studierenden als auch gegebenenfalls ihrer Ehepartner und Kinder abdeckt. Außerdem helfen wir unseren Stipendiaten bei der Wohnungssuche und organisatorischen Problemen aller Art, bieten Englisch- und Deutschkurse auf mehreren Niveaus, gemeinsame Aktivitäten und Exkursionen an. :::

KONTAKT

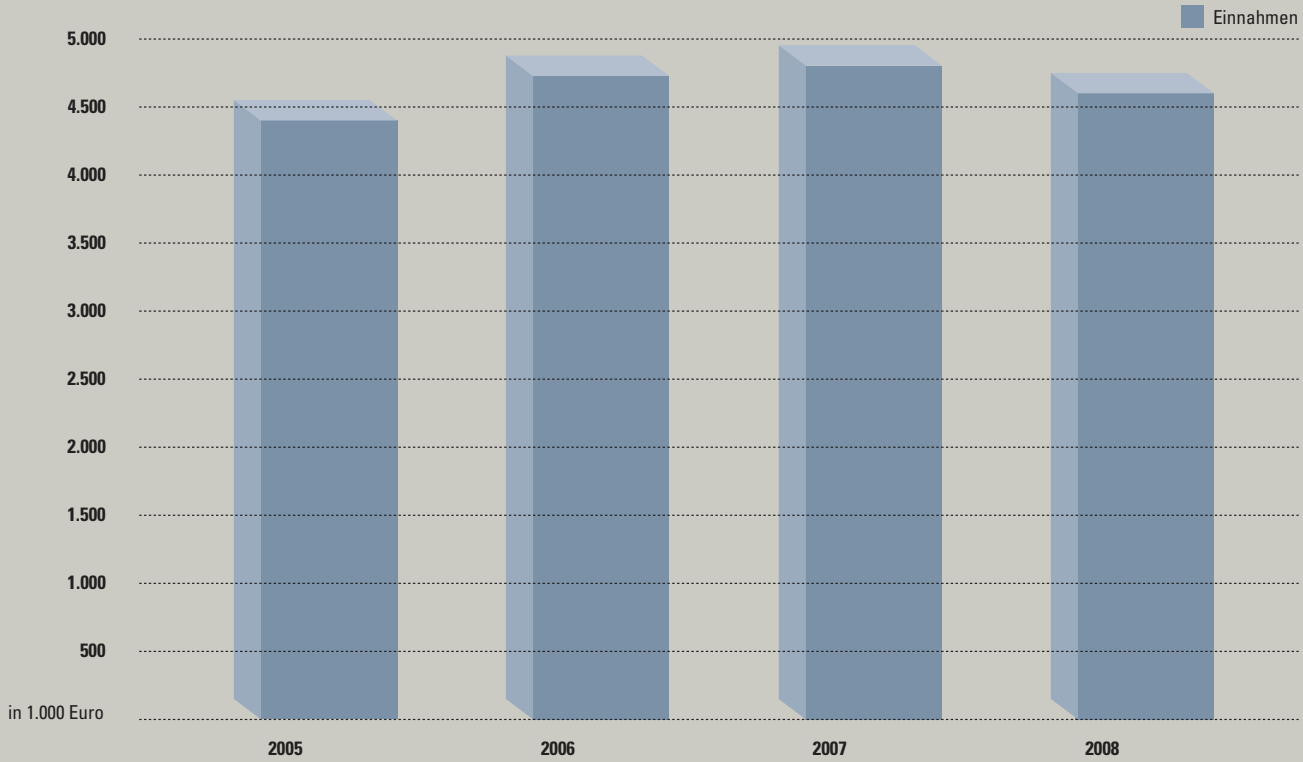
Jennifer Gerling
IMPRS-CS
Telefon +49 681 9325-226
Email jgerling@mpi-inf.mpg.de
Internet http://www.imprs-cs.de

Das Institut in Zahlen

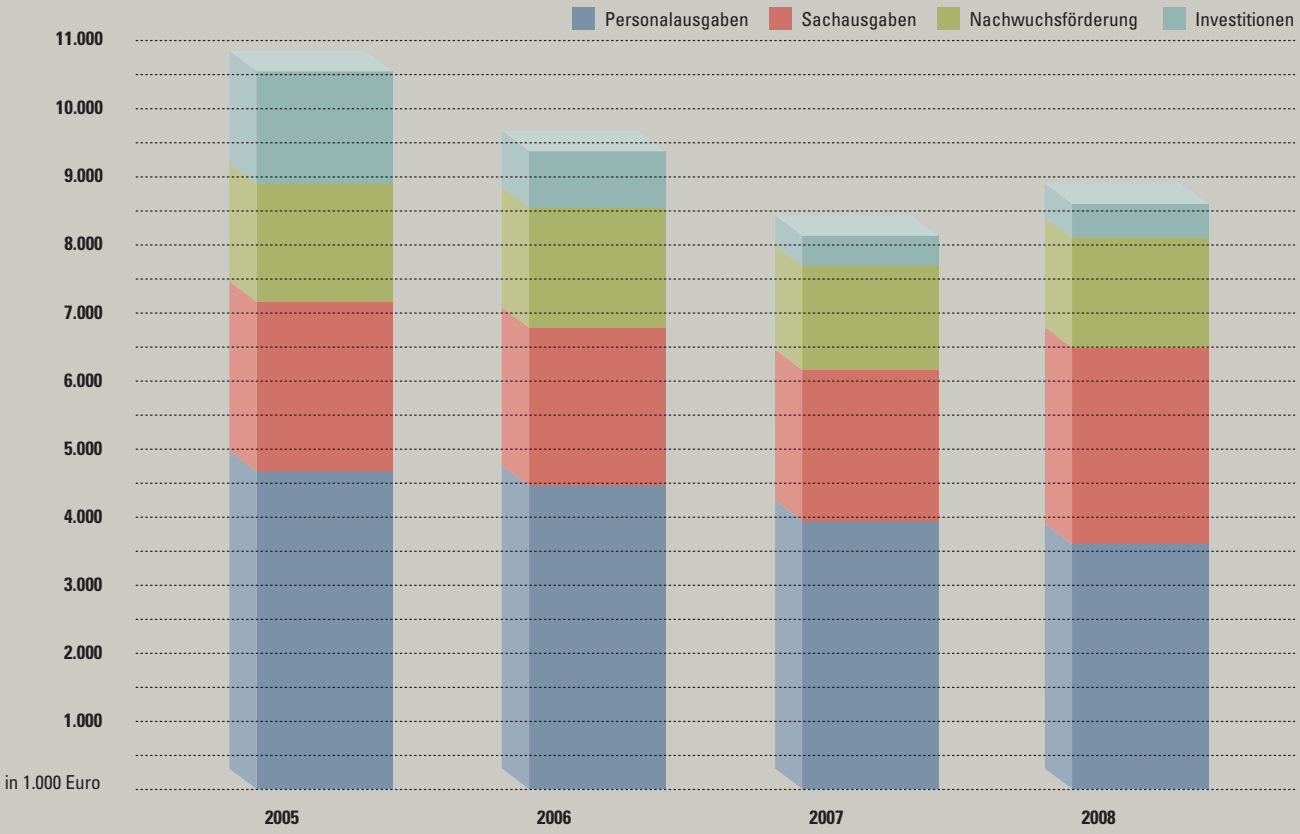
Drittmittelprojekte 2005 bis 2008 Anzahl und Verteilung



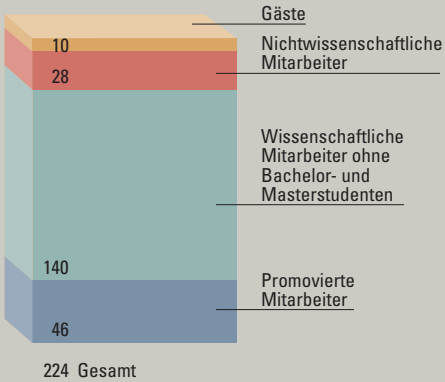
Drittmittelprojekte 2005 bis 2008



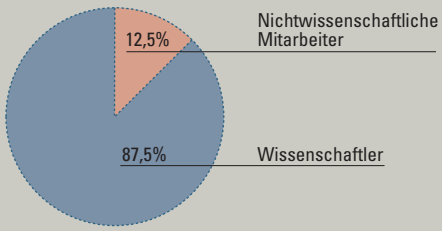
Betriebsmittel 2005 bis 2008



Mitarbeiter am Institut Stand 1.1.2009



Verhältnis von Wissenschaftlern zu nichtwissenschaftlichen Mitarbeitern am Institut Stand 1.1.2009



KONTAKT

Volker Geiß
Gemeinsame Verwaltung
Telefon +49 681 9325-700
Email geiss@mpi-inf.mpg.de

Rechnerbetrieb

Ungehinderte weltweite Kooperation und Kommunikation in einem motivierenden Umfeld bilden die Basis für ein Institut mit dem Anspruch, erstklassige Forschungsergebnisse hervorzubringen. Flexibilität, Qualität und Zuverlässigkeit der Ausstattung sowie ihre einfache Nutzbarkeit leisten dazu einen entscheidenden Beitrag.

Dieser Anspruch lässt sich auf unsere Rechnerinfrastruktur übertragen: Wir betreiben ein vielfältiges, sich schnell änderndes System, das sich dem Anwender einheitlich und verlässlich präsentiert, und das trotz der notwendigen Offenheit zur Unterstützung internationaler Kooperationen die Sicherheit nicht vernachlässigt.

Vielfalt der Werkzeuge

Wir setzen Systeme unterschiedlicher Hersteller ein. Die Auswahl wird durch die Anforderungen der Projekte bestimmt. Die Verfügbarkeit und preisliche Attraktivität von 64-Bit-Systemen auf der Basis von AMD oder Intel CPUs hat andere Systeme weitgehend verdrängt. Bei den Betriebssystemen verhält es sich ähnlich: Linux und Windows dominieren sowohl den Notebook- und Workstation- als auch den Server-Einsatz. Lediglich bei File-Servern setzen wir aufgrund von Stabilität und anderen Leistungsmerkmalen Solaris ein.

Dynamik und Innovation

Forschung an vorderster Front heißt auch, immer wieder innovative Technologien einzusetzen. Vor allem in der Computer-Grafik implizieren die möglichen Performance-Steigerungen und die Erweiterung des Funktionsumfangs neuester Hardwarekomponenten nicht selten die Notwendigkeit des Einsatzes von Prototypen.

Weitgehend einheitliche Nutzung

Die Forschungsprojekte sind häufig plattformübergreifend angelegt, da sie entweder für einen heterogenen Verbund gedacht sind, oder aber die Vorzüge der verschiedenen Rechner- und Betriebssystemarchitekturen ausnutzen müssen.

Mit Ausnahme weniger Spezialsysteme und der persönlich genutzten Notebooks sind daher alle Maschinen für jeden unserer weit über 900 Anwender (Mitarbeiter, Studenten, Projektpartner) ohne besonderen Aufwand direkt nutzbar – die User-Daten und die wichtigsten Softwarepakete sind plattformunabhängig verfügbar. Diese Homogenität erleichtert den Umgang mit dem Gesamtsystem erheblich und fördert seine Akzeptanz.

Zuverlässigkeit der Installationen

Ständige Updates und Upgrades bei Soft- und Hardware in Verbindung mit dem Wunsch, plattformübergreifend zu arbeiten, stellen hohe Anforderungen an die Zuverlässigkeit der Installations- und Administrationsvorgänge. Sie müssen reproduzierbar und verlässlich sein. Nach einer Umstellung oder Neuinstallation eines Systems sollte niemand gezwungen sein, seine Experimente oder gar seine Arbeitsweise anzupassen.

Kooperation und Kommunikation

Unser Netzwerk ist nach organisatorischen und sicherheitsrelevanten Gesichtspunkten in verschiedene Bereiche aufgeteilt. Die Endgeräte in den Teilnetzen der einzelnen Bereiche werden über ihren Etagen-Switch mit 10/100/1000-Mbit-Ethernet versorgt. Die Switche sind ausfallsicher mit 10-Gigabit-Ethernet an die Zentrale (zwei redundante Backbone-Switche) angeschlossen, die auch zentrale Server mit dieser Bandbreite versorgt. Server-Farmen und Compute-Cluster sind über eigene Switche redundant mit der Zentrale verbunden.

Der externe Bereich umfasst die notwendigen 1- und 10-GigaBit-Verbindungen zu verschiedenen Einrichtungen auf dem Campus der Universität des Saarlandes, in Saarbrücken und auf dem Universitätscampus in Kaiserslautern. Die Internet-Konnektivität wird über einen mit der Universität gemeinsam genutzten 2,6 Gigabit-XWIN-Anschluss des DFN-Vereins realisiert. Auch unsere externen Anschlüsse sind in der Regel doppelt ausgeführt, um die notwendige Ausfallsicherheit zu implementieren.

Extern zugreifbare aber anonyme Dienste (DNS (Internet-Adressbuch), WWW, FTP (Datentransfer), SMTP (E-Mail) etc.) werden an der Firewall des Institutes in mehreren entmilitarisierten Zonen (DMZs) zusammengefasst, die entsprechend ihrer Bedeutung und ihres Gefährdungspotentials unterschieden werden.

Organisatorisch ist das Netzwerk so strukturiert, dass die Integration von Gastwissenschaftlern und Studenten durch die Möglichkeit gefördert wird, mitgebrachte Notebooks ohne zusätzliche Softwareinstallation anschließen oder über WLAN betreiben zu können. Sie werden unabhängig von ihrem Standort in speziell dafür vorgesehene Teilnetze gelenkt. Auch diese Netze werden in einer entmilitarisierten Zone zusammengefasst, so dass z.B. virenbefallene Systeme nur begrenzten Schaden anrichten können.

Die internationale Kooperation verlangt externe Zugriffsmöglichkeiten auf interne Ressourcen der Infrastruktur (Intranet). Hier bieten wir unter anderem einen gesicherten Terminal-Zugang und den Zugriff auf E-Mail und andere wichtige Datenbanken und Dienste an. Die Kooperation bei der Softwareentwicklung wird durch einen geschützten Zugang zu einer Versionsdatenbank unterstützt (Software-Repository: Subversion).

Sicherheit und Schutz

Pauschale Schutzmechanismen gegen Sabotage und Spionage sind in offenen Systemen nicht möglich. Sie schränken ihre Nutzbarkeit zu sehr ein. Die Sicherheitsrichtlinien können deshalb nur ein Kompromiss sein, der flexibel den Anforderungen folgt.

Zwar können die Strukturierung des Netzwerks, die Firewall, die Verschlüsselung für externe Zugriffe oder der Virenscanner im Mail-Server einige direkte Gefahren abwenden. Indirekte Gefahren, die unter anderem durch den Anschluss virenverseuchter Rechner im Intranet oder durch Fehler in extern agierenden Software-Systemen entstehen, können nur durch aktuelle Softwarestände und kontinuierlich aktualisierte Virenscanner auf den Maschinen verhindert werden.

Automatisierung als Garantie

Betrachtet man die zuvor beschriebenen Merkmale der Infrastruktur, so ist ihre ständige Weiterentwicklung und Anpassung an die neuesten Anforderungen der Forschungsprojekte und des Sicherheitskonzeptes unter Beibehaltung der Homogenität und Verlässlichkeit unsere zentrale Aufgabe.

Bei der Vielzahl der eingesetzten Hard- und Softwarekomponenten, sind es die weitgehend von uns konzipierten und weiterentwickelten automatisierten Abläufe, die es uns ermöglichen, die notwendigen Arbeiten in angemessener Zeit auf den betroffenen Systemen zu erledigen.

In einer einheitlichen Hardwareumgebung ließe sich die Automatisierung in vielen Fällen durch Vervielfältigung besonders gepflegter Installationen durchführen (Image-Erstellung). Unser Umfeld dagegen ist zu dynamisch und zu heterogen für diese Vorgehensweise.

Aus diesem Grund favorisieren wir Betriebssysteme, deren Installationsmechanismen paketorientiert sind. Solaris, aber vor allem Debian im Linux-Umfeld setzen auf Paketsysteme, die Abhängigkeiten zwischen den installierten Paketen berücksichtigen. Für die Windows-Plattformen haben wir ein solches Paketsystem mit Hilfe zusätzlicher Software eingeführt (netinstall).

Dem Beispiel von Solaris folgend, haben wir für die aktuellen Betriebssystemversionen von Debian und Windows ein Installations- und Administrationssystem implementiert. Durch die Wartung und Weiterentwicklung dieses Systems administrieren wir automatisch alle betreuten Systeme und steuern Neuinstallationen. Einmal erzielte Ergebnisse sind beliebig wiederholbar und sehr schnell auf die ganze Infrastruktur anzuwenden. Diese Implementierungsarbeit kostet allerdings Zeit und verlangsamt in manchen Fällen die Reaktionszeiten. Die Vorteile für die Betriebssicherheit wiegen diesen Nachteil jedoch eindeutig auf.

Das System ist so flexibel gestaltet, dass auch Sonderfälle durch spezifische Erweiterungen schnell realisiert werden können. Nicht sinnvoll wäre allerdings die Integration kurzlebiger Spezialinstallationen.

Compute-Service

Neben Workstations, Notebooks und einigen kleineren Servern betreiben wir mehrere mid-size Systeme, die über 16 eng gekoppelte CPUs und 256 GB Hauptspeicher verfügen. Diese Maschinen werden für Applikationen genutzt, die eine hohe Parallelität und den uniformen Zugriff auf großen Hauptspeicher benötigen.

Unser bisher größtes, im Jahre 2005 in Betrieb genommenes, Dual-Opteron-Cluster (96 Systeme), wird Anfang 2010 durch ein neues Cluster mit 128 Systemen mit je 8 Kernen und 48 GByte Hauptspeicher ersetzt. Es wird, wie sein Vorgänger, unter der Sun Grid Engine (SGE) betrieben. Durch die automatische Verteilung der Prozesse auf die einzelnen Rechner des Clusters erreichen wir eine hohe Auslastung des Gesamtsystems. Durch die Homogenität der Betriebssysteminstallationen können die zuvor erwähnten mid-size Systeme in dieses System integriert werden und so zur Flexibilität der Infrastruktur beitragen.

Berechnungen werden vor ihrer Freigabe für das Cluster dem Bedarf entsprechend kategorisiert, so dass die SGE-Software die optimale Verteilung der für die Berechnung notwendigen Prozesse auf die verfügbare Hardware vornehmen kann. Die Priorisierung bestimmter Prozesse hilft bei der angemessenen und fairen Verteilung der Ressourcen auf die insgesamt anstehenden Aufgaben.

File-Service und Datensicherung

Die Daten des Instituts werden über mehrere File-Server per NFS und CIFS (SMB) zur Verfügung gestellt. Die mittlerweile ca. 77 TB zentralen Daten sind auf mehrere RAID-Systeme verteilt, die über ein redundant ausgelegtes Storage Area Network (SAN) angeschlossen sind. Alle RAID-Systeme werden paarweise gespiegelt betrieben, um auch gegen den Ausfall eines ganzen RAID-Systems geschützt zu sein.

Unsere Datensicherung basiert auf zwei unterschiedlichen Systemen: einem konventionellen Band-Backup, das die Daten direkt mittels Legato-Networker auf einen Band-Roboter sichert und einem Online-Disk-Backup-System, das mit Hilfe von Datenvergleichen den Platzbedarf minimiert (Open-Source-System BackupPC) und die gesicherten Daten immer Online bereit hält. Die Zahlen dieser Online-Datensicherung sind beeindruckend. Da keine Daten in diesem System wirklich doppelt gehalten werden, können ca. 615 Terabyte Brutto-Daten der verschiedenen Backupläufe auf etwa 36 Terabyte untergebracht werden. Um die Vorteile von Disk- und Tape-Technologien zu vereinen, werden wir dieses System in Zukunft mit dem Bandroboter kombinieren.

Das betagte Backup-System auf der Basis von SDLT-110/220-Laufwerken und der Legato-Software wird zur Zeit durch ein T10000B System von StorageTek und SUN-Microsystems mit IBM Tivoli Storage Manager ersetzt. Dieses System hat in der aktuellen Ausbaustufe die Möglichkeit auf 700 Bänder mit einer Gesamtkapazität von 700TB zuzugreifen. Es steht in einem eigens dafür vorbereiteten Raum, um dem Schutz der Datenkonserven schon bei ihrer Lagerung im Roboter einen maximalen Stellenwert einzuräumen.

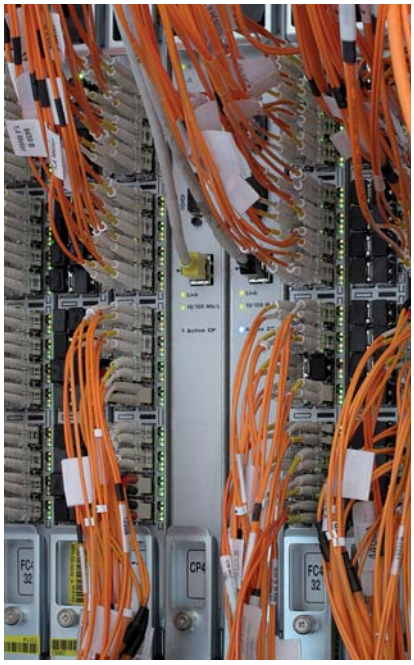
Spezialsysteme

Für spezielle Forschungsaufgaben, insbesondere aus dem Bereich der Computergrafik, werden diverse Spezialsysteme benötigt. Es stehen u.a. ein digitales Videoschnittsystem, mehrere 3D-Scanner, Multivideoaufnahmesysteme, Videokonferenzsysteme und 3D-Projektionssysteme zur Verfügung.

Betriebssicherheit

Die Maßnahmen zur Sicherheit (Firewall, VPN), die automatisierten Installations- und Konfigurationswerkzeuge, der Betrieb ausfallsicherer Disksysteme, der Betrieb eines ausfallsicheren Netzwerk-Backbones und die Datensicherung werden durch ein Überwachungssystem ergänzt, das über kritische Zustände der Serversysteme und des Netzwerks aber auch über Fehlfunktionen komplexer Prozesse per E-Mail und SMS informiert.

Stromversorgung und Kühlung sind so konzipiert, dass der Serverbetrieb auch bei Stromausfällen aufrechterhalten werden kann. Als letztes Glied in der Kette versorgt ein Generator mit einer Leistung von ca. einem Megawatt die Kerninfrastruktur und die Kühlungssysteme mit ausreichend Leistung, um einen ununterbrochenen Betrieb garantieren



zu können. Unser neuer Rechnerraum, der im Rahmen der Renovierung und Vergrößerung der Infrastruktur und ihrer Kapazitäten realisiert wurde, steht kurz vor der Inbetriebnahme. Damit werden wir dann in der Lage sein, die vielen Redundanzsysteme auch tatsächlich in zwei relativ weit voneinander entfernten und unabhängigen Räumen zu installieren und damit die Betriebssicherheit weiter zu erhöhen.

Zuständigkeiten

Einkauf, Installation, Administration, Betrieb, Anwenderbetreuung und Fortschreibung der beschriebenen Systeme und Techniken sind die Aufgaben von IST (Information Services and Technology), der inzwischen neu geschaffenen gemeinsamen IT-Abteilung der Institute für Informatik und Softwaresysteme. Bedingt durch das Engagement der beiden Institute in Kooperationen mit Fachbereichen und Instituten der Universität ist IST auch für die IT der gemeinsamen Campus-Bibliothek und des Exzellenz-Clusters „Multi Modal Computing and Interaction“ zuständig.

Flexible Unterstützung wissenschaftlicher Projekte

Die beschriebenen Dienste, Server und Compute-Cluster werden für eine Vielzahl wissenschaftlicher Projekte in den unterschiedlichsten Szenarien eingesetzt, teilweise auch in internationalen Kooperationen wie beispielsweise der „CADE ATP System Competition“ [http://www.cs.miami.edu/~tptp/CASC/22/]. Um den daraus resultierenden und sehr unterschiedlichen Anforderungen gerecht werden zu können, unterstützt IST eine fein abgestufte Betreuung, die vom

reinen Server-Hosting in Einzelfällen bis hin zur Applikationsbetreuung reicht. Die Aufgabentrennung ergibt sich in den meisten Fällen aus der Schnittmenge der Projektanforderungen und dem Portfolio von IST. Zur Lösung von Problemen mit selbstadministrierten Systemen steht IST dann allerdings nur noch beratend zur Verfügung.

Personalstruktur

Neben Leitung und Einkauf (zwei Stellen) steuert das Max-Planck-Institut für Informatik sieben wissenschaftliche Mitarbeiter und einen Techniker bei. IST wird ergänzt durch drei wissenschaftliche Mitarbeiter des Max-Planck-Institut für Softwaresysteme und einen Techniker. Für die zusätzlichen Administrationsaufgaben zur Unterstützung der gemeinsamen Bibliothek und des Exzellenzclusters steht jeweils ein Mitarbeiter zur Verfügung.

IST teilt sich in eine Core- und mehrere Front-Gruppen auf. Vereinfacht dargestellt ist die Core-Gruppe für

Dienste und Dienstleistungen zuständig, die für beide Institute identisch sind oder sogar gemeinsam betrieben werden. Die Front-Gruppen decken dementsprechend die spezifischen Bedürfnisse der Institute ab. Die Mitglieder der Front-Gruppen sind damit auch eher an den verschiedenen Standorten tätig.

Die direkte Anwenderbetreuung übernimmt ein achtköpfiges Team von Studenten. Sie bilden zusammen mit drei Jahres-Praktikanten die Kommunikationsschnittstelle für die Anwender und sind einerseits virtuell per Mail oder Web-Interface, zu Bürozeiten aber auch persönlich erreichbar. Neben der Bearbeitung von Fragen zur Benutzung der Infrastruktur pflegt diese Gruppe auch Informationssysteme, die u.a. eine Zusammenfassung der interessanten Fragen und Antworten (FAQ) und ein Bulletin-Board umfassen. :::

KONTAKT

Jörg Herrmann
IT-Abteilung
Telefon +49 681 9325-5800
Email jh@mpi-inf.mpg.de
Internet http://www.mpi-inf.mpg.de/services/ist/

Ausgewählte Kooperationen

BIOINFORMATIK

::: BioSolveIT GmbH, *Sankt Augustin, Deutschland*

::: Centrum Wiskunde & Informatica, *Amsterdam, Niederlande*

::: Christian-Albrechts-Universität, *Kiel, Deutschland*

::: CSIRO Livestock Industries, *St. Lucia, Queensland, Australien*

::: European Bioinformatics Institute, *Hinxton, Großbritannien*

::: Harvard University and Broad Institute, Cambridge, *Massachusetts, USA*

::: Institute for Clinical Molecular Biology, *Kiel, Deutschland*

::: Johann-Wolfgang-Goethe-Universität Frankfurt, *Frankfurt, Deutschland*

::: Karolinska Institut, *Stockholm, Schweden*

::: Labor Dr. Thiele, *Kaiserslautern, Deutschland*

::: Max-Delbrück-Centrum für Molekulare Medizin, *Berlin, Deutschland*

::: Max-Planck-Institut für Neurologische Forschung, *Köln, Deutschland*

::: Max-Planck-Institut für Molekulare Physiologie, *Dortmund, Deutschland*

::: Max-Planck-Institut für Biophysikalische Chemie, *Göttingen, Deutschland*

::: Ruprecht-Karls-Universität, *Heidelberg, Deutschland*

::: Swammerdam Institute for Life Sciences, *Amsterdam, Niederlande*

::: Technische Universität Dortmund, *Dortmund, Deutschland*

::: Università di Siena, *Siena, Italien*

::: Universität Düsseldorf, *Düsseldorf, Deutschland*

::: Universität Heidelberg, *Heidelberg, Deutschland*

::: Universität Köln, *Köln, Deutschland*

::: Universität des Saarlandes, *Saarbrücken, Deutschland*

::: Universität Salzburg, *Salzburg, Österreich*

::: University of Barcelona, *Barcelona, Spanien*

::: University of British Columbia, *Vancouver, Kanada*

::: University of California, Santa Cruz, *California, USA*

::: University of Kansas, *Kansas, USA*

::: Wellcome Trust Sanger Institute, *Hinxton, Großbritannien*

GEOMETRIE

::: Eidgenössische Technische Hochschule Zürich, *Zürich, Schweiz*

::: Freie Universität Berlin, *Berlin, Deutschland*

::: GeometryFactory, *Grasse, Frankreich*

::: INRIA, *Sophia-Antipolis, Frankreich*

::: Microsoft Research, *Cambridge, Großbritannien*

::: National and Kapodistrian Univerity of Athens, *Athen, Griechenland*

::: Stanford University, *Stanford, USA*

::: Tel Aviv University, *Tel Aviv, Israel*

::: Universität Hannover, *Hannover, Deutschland*

::: Universität des Saarlandes, *Saarbrücken, Deutschland*

::: Universität Tübingen, *Tübingen, Deutschland*

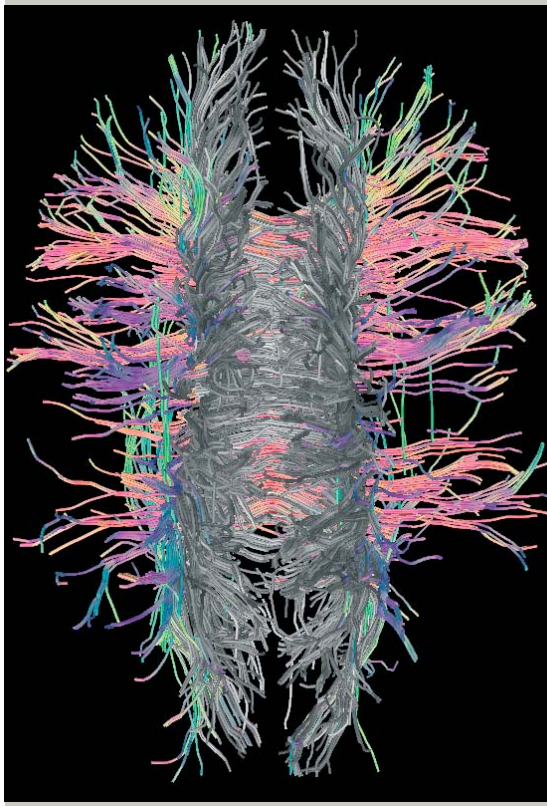
::: University of Adelaide, *Adelaide, Australien*

::: University of Groningen, *Groningen, Niederlande*

GARANTIEN

::: Indian Institute of Technology, *Neu Delhi, Indien*

::: Laboratoire Lorrain de Recherche en Informatique et ses Applications, *Nancy, Frankreich*



::: Max-Planck-Institut für Softwaresysteme, *Saarbrücken, Deutschland*

::: National Information and Communications Technology Australia, *Canberra, Australien*

::: Università di Milano, *Mailand, Italien*

::: Universität Freiburg, *Freiburg, Deutschland*

::: Universität Oldenburg, *Oldenburg, Deutschland*

::: University of Liverpool, *Liverpool, Großbritannien*

INFORMATIONSSUCHE & DIGITALES WISSEN

::: Consiglio Nazionale delle Ricerche (CNR), *Pisa, Italien*

::: Eidgenössische Technische Hochschule Zürich, *Zürich, Schweiz*

::: European Archive Foundation, *Amsterdam, Niederlande*

::: Google, *Zürich, Schweiz*

::: Hebrew University, *Jerusalem, Israel*

::: IBM, *Haifa, Israel*

::: International Computer Science Institute Berkeley, *Berkeley, USA*

::: L3S Research Center, *Hannover, Deutschland*

::: Masaryk University of Brno, *Brno, Tschechien*

::: Microsoft Research, *Cambridge, Großbritannien*

::: Microsoft Research, *Redmond, USA*

::: New York University, *New York, USA*

::: Rényi Institute of the Hungarian Academy of Sciences, *Budapest, Ungarn*

::: Simon Fraser University, *Kanada*

::: The European Library, *The Hague, Niederlande*

::: Università di Padova, *Padua, Italien*

::: Universität Basel, *Basel, Schweiz*

::: Universität des Saarlandes, *Saarbrücken, Deutschland*

::: Universität Duisburg-Essen, *Duisburg, Deutschland*

::: University of Athens, *Athen, Griechenland*

::: University of Hong Kong, *Hong Kong, China*

::: University of Massachusetts, *Lowell, USA*

::: University of South Carolina, *Columbia, USA*

::: Yahoo! Research, *Barcelona, Spanien*

OPTIMIERUNG

::: Eidgenössische Technische Hochschule Zürich, *Zürich, Schweiz*

::: Indian Institute of Technology, *Kanpur, Indien*

::: INRIA Nancy - Grand Est Research Centre, *Nancy, Frankreich*

::: Maastricht University, *Maastricht, Niederlande*

::: MIT Computer Science and Artificial Intelligence Lab, *Cambridge, USA*

::: New York University, *New York, USA*

::: Sobolev Institute of Mathematics, *Omsk, Russland*

::: Technical University of Denmark, *Kopenhagen, Dänemark*

::: Technische Universität Berlin, *Berlin, Deutschland*

::: Technische Universität Dortmund, *Dortmund, Deutschland*

::: Technische Universität Wien, *Wien, Österreich*

::: Università di Roma „La Sapienza“, *Rom, Italien*

::: Universidad de Chile, *Santiago, Chile*

::: Universität Freiburg, *Freiburg, Deutschland*

::: Universität Kiel, *Kiel, Deutschland*

::: University of Adelaide, *Adelaide, Australien*

::: University of Birmingham, *Birmingham, Großbritannien*

::: Vrije Universiteit Amsterdam, *Amsterdam, Niederlande*

::: Warwick Mathematics Institute, *Warwick, Großbritannien*

SOFTWARE

::: Forschungsgesellschaft für Angewandte Naturwissenschaften, *Wachtberg, Deutschland*

::: Hochschule für Musik Saar, *Saarbrücken, Deutschland*

::: Linköping University, *Linköping, Schweden*

::: Los Alamos National Laboratory, *Los Alamos, USA*

::: PROSTEP AG, *Darmstadt, Deutschland*

::: Rheinische Friedrich-Wilhelms-Universität Bonn, *Bonn, Deutschland*

::: Universität Dortmund, *Dortmund, Deutschland*

::: Universität Hannover, *Hannover, Deutschland*

::: Universität Siegen, *Siegen, Deutschland*

::: Universtity of California, *Davis, USA*

VISUALISIERUNG

::: Adobe System Inc., *San Jose, USA*

::: INRIA Rhône-Alpes, *Grenoble, Frankreich*

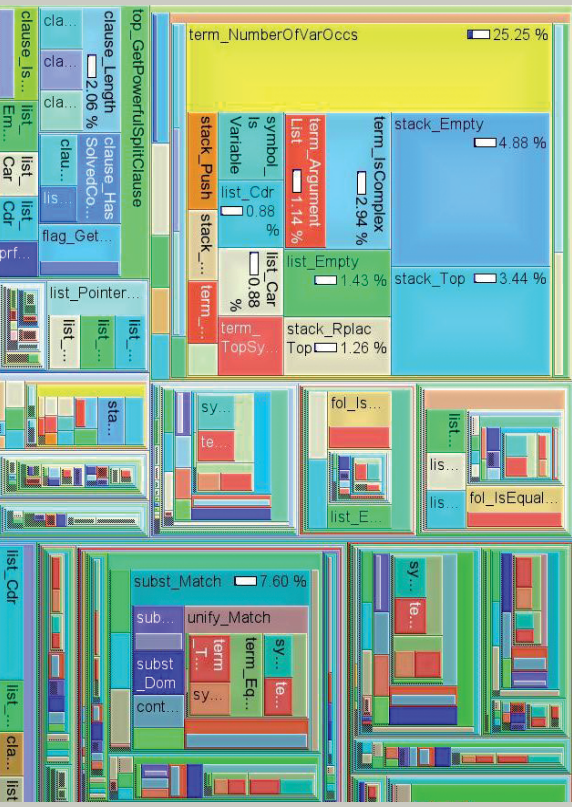
::: INRIA, Sophia-Antipolis, *Frankreich*

::: Massachusetts Institute of Technology, *Cambridge, USA*

::: Max-Planck-Institut für Kognitions- und Neurowissenschaften, *Leipzig, Deutschland*

::: University of Illinois, *Urbana, USA*

Ausgewählte Publikationen



[1] E. de Aguiar, C. Stoll, C. Theobalt, N. Ahmed, H.-P. Seidel and S. Thrun. Performance capture from sparse multi-view video. *ACM Trans. Graph.*, 27(3):98ff, 2008.

[2] N. Ahmed, C. Theobalt, C. Rössl, S. Thrun and H.-P. Seidel. Dense correspondence finding for parametrization-free animation reconstruction from video. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2008)*, Anchorage, Alaska, 2008, pp. 1–8. IEEE Computer Society.

[3] W. Allasia, F. Falchi, F. Gallo, M. Kacimi, A. Kaplan, J. Mamou, Y. Mass and N. Orio. Audiovisual content analysis in P2P networks: The SAPIR approach. In *The 19th International Conference on Database and Expert Systems Application (DEXA 2008)*, Turin, Italy, August 2008, pp. 610–614. IEEE Computer Society.

[4] E. Althaus, E. Kruglov and C. Weidenbach. Superposition modulo linear arithmetic sup(la). In S. Ghilardi and R. Sebastiani, eds., *7th international Symposium on Frontiers of Combining Systems (FroCos 2009)*, Trento, Italy, September 2009, LNAI 5749, pp. 84–99. Springer.

[5] A. Altmann, M. D aumer, N. Beerenwinkel, Y. Peres, E. Sch ulter, J. B uch, S.-Y. Rhee, A. S onnerborg, W. J. Fessel, R. W. Shafer, M. Zazzi, R. Kaiser and T. Lengauer. Predicting the response to combination antiretroviral therapy: Retrospective validation of geno2pheno-THEO on a large clinical database. *The Journal of Infectious Diseases*, 199:999–1006, April 2009.

[6] A. Altmann, T. Sing, H. Vermeiren, B. Winters, E. Van Craenenbroeck, K. Van der Borght, S.-Y. Rhee, R. W. Shafer, E. Sch ulter, R. Kaiser, Y. Peres, A. S onnerborg, W. J. Fessel, F. Incardona, M. Zazzi, L. Bacheler, H. Van Vlijmen and T. Lengauer. Advantages of predicted phenotypes and statistical learning models in inferring virological response to antiretroviral therapy from HIV genotype. *Antiviral Therapy*, 14(2):273–283, 2009.

[7] A. Anand, S. Bedathur, K. Berberich, R. Schenkel and C. Tryfonopoulos. EverLast: A distributed architecture for preserving the Web. In *Proceedings of the Joint Conference on Digital Libraries (JCDL 2009)*, Austin, Texas, 2009.

[8] S. Angelopoulos and P. Schweitzer. Paging and list update under bijective analysis. In C. Mathieu, ed., *20th ACM-SIAM Symposium on Discrete Algorithms (SODA 2009)*, New York, 2009, pp. 1136–1145. ACM press.

[9] T. Annen, Z. Dong, T. Mertens, P. Bekaert, H.-P. Seidel and J. Kautz. Real-time, all-frequency shadows in dynamic scenes. *ACM Trans. Graph.*, 27(3):1–8, 2008.

[10] I. Arikan, S. Bedathur and K. Berberich. Time Will Tell: Leveraging temporal expressions in IR. In *Second ACM International Conference on Web Search and Data Mining (WSDM 09) – Late Breaking Results*, Barcelona, Spain, 2009. ACM.

[11] Y. Assenov, F. Ramírez, S.-E. Schelhorn, T. Lengauer and M. Albrecht. Computing topological parameters of biological networks. *Bioinformatics*, 24(2):282–284, 2008.

[12] T. O. Aydin, R. Mantiuk, K. Myszkowski and H.-P. Seidel. Dynamic range independent image quality assessment. In G. Turk, ed., *Proceedings of ACM SIGGRAPH 2008*, Los Angeles, USA, 2008, *ACM Transactions on Graphics*, vol. 27(3), pp. 1–10. ACM.

[13] A. Baak, B. Rosenhahn, M. Müller and H.-P. Seidel. Stabilizing motion tracking using retrieved motion priors. In *IEEE 12th International Conference on Computer Vision (ICCV)*, sep 2009, pp. 1428–1435.

[14] N. Bansal and H.-L. Chan. Weighted flow time does not admit O(1)-competitive algorithms. In C. Mathieu, ed., *Proceedings of the Twentieth Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, New York, USA, 2009, pp. 1238–1244. ACM Press.

[15] H. Bast, C. W. Mortensen and I. Weber. Output-sensitive autocompletion search. *Inf. Retr.*, 11(4):269–286, 2008.

[16] P. Baumgartner and U. Waldmann. Superposition and model evolution combined. In R. A. Schmidt, ed., *Automated Deduction, CADE-22, 22nd International Conference on Automated Deduction*, Montreal, Canada, August 2009, LNAI 5663, pp. 17–34. Springer.

[17] E. Berberich, M. Kerber and M. Sagraloff. An efficient algorithm for the stratification and triangulation of an algebraic surface. *Computational Geometry: Theory and Applications (CGTA)*, 43(3):257–278, 2009.

[18] E. Berberich and M. Sagraloff. A generic and flexible framework for the geometrical and topological analysis of (algebraic) surfaces. *Comput. Aided Geom. Des.*, 26(6):627–647, 2009.

[19] K. Berberich, S. Bedathur, T. Neumann and G. Weikum. FluxCapacitor: Efficient time-travel text search. In C. L. Williamson, M. E. Zurko and P. J. Patel-Schneider, Peter F. Shenoy, eds., *33rd International Conference on Very Large Databases (VLDB 2007)*, Vienna, Austria, 2007, pp. 1414–1417. ACM.

[20] K. Berberich, S. Bedathur, T. Neumann and G. Weikum. A time machine for text search. In W. Kraaij, A. P. de Vries, C. L. A. Clarke, N. Fuhr and N. Kando, eds., *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2007)*, Amsterdam, Netherlands, 2007, pp. 519 –526. ACM.

[21] K. Berberich, S. Bedathur and G. Weikum. Tunable word-level index compression for versioned corpora. In *Proceedings of the*

Efficiency Issues in Information Retrieval Workshop, co-located with ECIR 2008, Glasgow, Scotland, 2008, Lecture Notes in Computer Science. Springer.

[22] H. Blankenburg, R. D. Finn, A. Prlić, A. M. Jenkinson, F. Ramírez, D. Emig, S.-E. Schelhorn, B. Joachim, T. Lengauer and M. Albrecht. Dasmi: exchanging, annotating and assessing molecular interaction data. *Bioinformatics*, 25(10):1321–1328, 2009.

[23] C. Bock, K. Halachev, J. Büch and T. Lengauer. EpiGRAPH: user-friendly software for statistical analysis and prediction of (epi)-genomic data. *Genome Biology*, 10:R14, 2009.

[24] C. Bock and T. Lengauer. Computational epigenetics. *Bioinformatics*, 24(1):1–10, 2008.

[25] C. Bock, J. Walter, M. Paulsen and T. Lengauer. Inter-individual variation of DNA methylation and its implications for large-scale epigenome mapping. *Nucleic Acids Research*, 36(10):e55, June 2008.

[26] M. Bokeloh, A. Berner, M. Wand, H.-P. Seidel and A. Schilling. Symmetry detection using line features. *Computer Graphics Forum (Proc. EUROGRAPHICS)*, 28(2):697–706, 2009.

[27] M. Celikik and H. Bast. Fast error-tolerant search on very large texts. In D. Shin, ed., *The 24th Annual ACM Symposium on Applied Computing*, Honolulu, Hawaii, USA, March 2009, *PROCEEDINGS OF THE 2009 ACM SYMPOSIUM ON APPLIED COMPUTING*, vol. 104092, pp. 1724–1731. ACM.

[28] C. Crassin, F. Neyret, S. Lefebvre and E. Eisemann. Gigavoxels : Ray-guided streaming for efficient and detailed voxel rendering. In *ACM SIGGRAPH Symposium on Interactive 3D Graphics and Games (I3D)*, Boston, MA, Etats-Unis, feb 2009. ACM Press. to appear.

[29] W. Damm, S. Disch, H. Hungar, S. Jacobs, J. Pang, F. Pigorsch, C. Scholl, U. Waldmann and B. Wirtz. Exact state set representations in the verification of linear hybrid systems with large discrete state space. In K. S. Namjoshi, T. Yoneda, T. Higashino and Y. Okamura, eds., *Automated Technology for Verification and Analysis, 5th International Symposium, ATVA 2007*, Tokyo, Japan, 2007, LNCS 4762, pp. 425–440. Springer.

[30] B. Doerr, T. Friedrich and T. Sauerwald. Quasirandom rumor spreading. In *Proceedings of the 19th Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2008*, San Francisco, USA, 2008, pp. 773–781. ACM.

[31] B. Doerr, T. Friedrich and T. Sauerwald. Quasirandom rumor spreading: Expanders, push vs. pull, and robustness. In *Proceedings of the 36th International Colloquium on Automata, Languages and Programming (ICALP 2009)*, Rhodes, Greece, 2009. Springer.

[32] B. Doerr, E. Happ and C. Klein. Crossover can provably be useful in evolutionary computation. In C. Ryan and M. Keijzer, eds., *Genetic and Evolutionary Computation Conference 2008*, Atlanta, USA, 2008, Proceedings of the 10th annual conference on Genetic and evolutionary computation, pp. 539–546. ACM. Best paper award.

[33] A. Eigenwillig and M. Kerber. Exact and efficient 2d-arrangements of arbitrary algebraic curves. In *SODA*, 2008, pp. 122–131.

[34] E. Eisemann, S. Paris and F. Durand. A visibility algorithm for converting 3d meshes into editable 2d vector graphics. *ACM Trans. Graph.*, 28(3), 2009.

[35] F. Eisenbrand, A. Karrenbauer, M. Skutella and C. Xu. Multiline addressing by network flow. *Algorithmica*, 53(4):583–596, April 2009.

[36] D. Emig, M. S. Cline, K. Klein, A. Kunert, P. Mutzel, T. Lengauer and M. Albrecht. Integrative visual analysis of the effects of alternative splicing on protein domain interaction networks. *Journal of Integrative Bioinformatics*, 5(2):101–115, 2008.

[37] D. Emig, M. S. Cline, T. Lengauer and M. Albrecht. Integrating expression data with domain interaction networks. *Bioinformatics*, 24(21):2546–2548, 2008.

[38] L. Epstein and R. van Stee. The price of anarchy on uniformly related machines revisited. In *1st Symposium on Algorithmic Game Theory (SAGT 2008)*, 2008, pp. 46–57. Springer.

[39] F. Falchi, M. Kacimi, Y. Mass, F. Rabitti and P. Zezula. SAPIR: Scalable and distributed image searching. In *SAMT (Posters and Demos)*, Genoa, Italy, November 2007, *CEUR Workshop Proceedings*, vol. 300, pp. 11–12. CEUR-WS.org.

[40] A. Franke, T. Balschun, T. H. Karlsen, J. Sventoraityte, S. Nikolaus, G. Mayr, F. S. Domingues, M. Albrecht, M. Nothnagel, D. Ellinghaus, C. Sina, C. M. Onnie, R. K. Weersma, P. C. F. Stokkers, C. Wijmenga, M. Gazouli, D. Strachan, W. L. McArdle, S. Vermeire, P. Rutgeers, P. Rosenstiel, M. Krawczak, M. H. Vatn, the IBSEN study group, C. G. Mathew and S. Schreiber. Sequence variants in IL10, ARPC2, and multiple other loci contribute to ulcerative colitis. *Nature Genetics*, 40(11):1319–1323, 2008.

[41] T. Friedrich, C. Horoba and F. Neumann. Multiplicative approximations and the hypervolume indicator. In G. Raidl and F. Rothlauf, eds., *Genetic and Evolutionary Computation Conference 2009*, Montreal, Canada, 2009, pp. 571–578. ACM. Best Paper Award in the track „Evolutionary Multiobjective Optimization".

[42] T. Friedrich and T. Sauerwald. Near-perfect load balancing by randomized rounding. In *Proceedings of the 41st ACM Symposium on Theory of Computing (STOC 2009)*, Bethesda, Maryland, USA, 2009. ACM.

[43] K. H. Gartemann, B. Abt, T. Bekel, A. Burger, J. Engemann, M. Flügel, L. Gaigalat, A. Goesmann, I. Gräfen, J. Kalinowski, O. Kaup, O. Kirchner, L. Krause, B. Linke, A. McHardy, F. Meyer, S. Pohle, C. Rückert, S. Schneiker, E. M. Zellermann, A. Pühler, R. Eichenlaub, O. Kaiser and D. Bartels. The genome sequence of the tomato-pathogenic actinomycete clavibacter michiganensis subsp. michiganensis ncppb382 reveals a large island involved in pathogenicity. *The Journal of Bacteriology*, 190(6):2138–2149, March 2008.

[44] D. Göddeke, R. Strzodka, J. Mohd-Yusof, P. McCormick, H. Wobker, C. Becker and S. Turek. Using GPUs to improve multigrid solver performance on a cluster. *International Journal of Computational Science and Engineering (IJCSE)*, 4(1):36–55, 2008.

[45] D. Göddeke, H. Wobker, R. Strzodka, J. Mohd-Yusof, P. McCormick and S. Turek. Co-processor acceleration of an unmodified parallel solid mechanics code with FEASTGPU. *International Journal of Computational Science and Engineering (IJCSE)*, 4(4), 2009.

[46] R. Harren and R. van Stee. Improved absolute approximation ratios for two-dimensional packing problems. In *12th Intl. Workshop on Approximation Algorithms for Combinatorial Optimization Problems (APPROX 2009)*, 2009, pp. 177–189.

[47] C. Hartmann, I. Antes and T. Lengauer. Docking and scoring with alternative side-chain conformations. *Proteins: Structure, Function, and Bioinformatics*, 74(3):712–726, February 2009.

[48] N. Hasler, B. Rosenhahn, T. Thorm ahlen, M. Wand, J. Gall and H.-P. Seidel. Markerless motion capture with unsynchronized moving cameras. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009)*, Miami, USA, June 2009. IEEE Computer Society.

[49] R. Herzog, K. Myszkowski and P. Szczepanski. Anisotropic radiance-cache splatting for efficiently computing high-quality global illumination with lightcuts. *Computer Graphics Forum (Proc. of EUROGRAPHICS)*, 28(3):259 – 268, 2009.

[50] L. Z. Holland, R. Albalat, K. Azumi, E. Benito-Gutiérrez, M. J. Blow, M. Bronner-Fraser, F. Brunet, T. Butts, S. Candiani, L. J. Dishaw, D. E. Ferrier, J. Garcia-Fernàndez, J. J. GibsonBrown, C. Gissi, A. Godzik, F. Hallböök, D. Hirose, K. Hosomichi, T. Ikuta, H. Inoko, M. Kasahara, J. Kasamatsu, T. Kawashima, A. Kimura, M. Kobayashi, Z. Kozmik, K. Kubokawa, V. Laudet, G. W. Litman, A. McHardy, D. Meulemans, M. Nonaka, R. P. Olinski, Z. Pancer, L. A. Pennacchio, M. Pestarino, J. P. Rast, I. Rigoutsos, M. Robinson-Rechavi, G. Roch,

H. Saiga, Y. Sasakura, M. Satake, Y. Satou, M. Schubert, N. Sherwood, T. Shiina, N. Takatori, J. Tello, P. Vopalensky, S. Wada, A. Xu, Y. Ye, K. Yoshida, F. Yoshizaki, J. K. Yu, Q. Zhang, C. M. Zmasek, P. J. de Jong, K. Osoegawa, N. H. Putnam, D. S. Rokhsar, N. Satoh and P. W. Holland. The amphioxus genome illuminates vertebrate origins and cephalochordate biology. *Genome Research*, 18(7):1100–1111, July 2008.

[51] M. Horbach and C. Weidenbach. Superposition for fixed domains. In *CSL*, Bertinoro, Italy, 2008, LNCS 5213, pp. 293–307. Springer.

[52] C. Ihlemann, S. Jacobs and V. Sofronie-Stokkermans. On local reasoning in verification. In C. R. Ramakrishnan and J. Rehof, eds., *Proceedings of TACAS 2008*, Budapest, Hungary, 2008, LNCS 4963, pp. 265–281. Springer.

[53] C. Ihlemann and V. Sofronie-Stokkermans. System description: H-PILoT. In R. Schmidt, ed., *22nd International Conference on Automated Deduction (CADE-22)*, Montreal, Canada, August 2009, Lecture Notes in Artificial Intelligence. Springer.

[54] S. Jacobs. Incremental instance generation in local reasoning. In A. Bouajjani and O. Maler, eds., *Computer Aided Verification – 21st International Conference, CAV 2009*, Grenoble, France, 2009. Springer. To appear.

[55] M. G. Kalyuzhnaya, A. Lapidus, N. Ivanova, A. C. Copeland, A. McHardy, E. Szeto, A. Salamov, I. V. Grigoriev, D. Suciu, S. R. Levine, V. M. Markowitz, I. Rigoutsos, S. G. Tringe, D. C. Bruce, P. M. Richardson, M. E. Lidstrom and L. Chistoserdova. High-resolution metagenomics targets specific functional types in complex microbial communities. *Nature Biotechnology*, 26(9):1029–34, September 2008.

[56] G. Kasneci, M. Ramanath, M. Sozio, F. Suchanek and G. Weikum. STAR: Steiner-tree approximation in relationship graphs. In *Proceedings of the 25th International Conference on Data Engineering (ICDE 2009)*, Shanghai, China, 2009. IEEE Computer Society.

[57] G. Kasneci, F. Suchanek, G. Ifrim, S. Elbassuoni, M. Ramanath and G. Weikum. NAGA: Harvesting, searching and ranking knowledge (demo). In J. Tsong-Li Wang, ed., *Proceedings of the ACM SIGMOD 2008 International Conference on Management of Data (SIGMOD 2008)*, Vancouver, Canada 2008, June 2008, pp. 1285–1288. ACM.

[58] G. Kasneci, F. Suchanek, G. Ifrim, M. Ramanath and G. Weikum. NAGA: Searching and ranking knowledge. In *24th International Conference on Data Engineering (ICDE 2008)*, Cancun, Mexico, 2008, pp. 953–962. IEEE Computer Society.

[59] S. Kosov, K. Scherbaum, K. Faber, T. Thormählen and H.-P. Seidel. Rapid stereo-vision enhanced face detection. In *IEEE International Conference on Image Processing (ICIP 2009)*, Cairo, Egypt, 2009, pp. 1221–1224. IEEE.

[60] S. Kratsch. Polynomial kernelizations for MIN F+Pi1 and MAX NP. In S. Albers and J.Y. Marion, eds., *26th International Symposium on Theoretical Aspects of Computer Science (STACS)*, Freiburg, Germany, March 2009. Schloss Dagstuhl – Leibniz-Zentrum für Informatik.

[61] S. Kratsch and F. Neumann. Fixed-parameter evolutionary algorithms and the vertex cover problem. In G. Raidl and F. Rothlauf, eds., *Genetic and Evolutionary Computation Conference 2009*, Montreal, Canada, 2009, pp. 293–300. ACM. Best Paper Award in the track „Combinatorial Optimization and Metaheuristics".

[62] F. Kurth and M. Müller. Efficient index-based audio matching. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(2):382–395, February 2008.

[63] M. Lamotte-Schubert and C. Weidenbach. Analysis of authorizations in SAP R/3. In *FTP 2009 Workshop Proceedings*, Oslo, Norway, July 2009, University Oslo Research Report, pp. 90–104. University of Oslo.

[64] R. Lasowski, A. Tevs, H.-P. Seidel and M. Wand. A probabilistic framework for partial intrinsic symmetries in geometric data. In *IEEE International Conference on Computer Vision (ICCV)*, September 2009.

[65] S. Lee, E. Eisemann and H.-P. Seidel. Depth-of-Field Rendering with Multiview Synthesis. *ACM Transactions on Graphics (Proc. ACM SIGGRAPH ASIA)*, 28(5):134, 2009.

[66] T. Lengauer and R. Kaiser. Computerjagd auf das AIDS virus. *Spektrum der Wissenschaft*, pp. 62–67, August 2009.

[67] A. McHardy. Finding genes in genome sequence. In J. M. Keith, ed., *Bioinformatics. - Vol. 1, Data, Sequence Analysis and Evolution, Methods in Molecular Biology*, vol. 452, ch. 2, pp. 163–77. Humana Press, Clifton, N.J., USA, 2008.

[68] A. McHardy and B. Adams. The role of genomics in tracking the evolution of influenza A virus. *PLoS Pathogens*, 5(10):e1000566, 2009.

[69] K. Mehlhorn and M. Sagraloff. Isolating real roots of real polynomials. In *ISSAC 09*, New York, NY, USA, 2009, pp. 247–254. ACM.

[70] J. Mestre. Adaptive local ratio. In *19th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, San Francisco, USA, January 2008, pp. 152–160. Society for Industrial and Applied Mathematics.

[71] U. Mihm, O. Ackermann, C. Welsch, E. Herrmann, W.-P. Hofmann, N. Grigorian, H. Welker, T. Lengauer, S. Zeuzem and C. Sarrazin. Clinical relevance of the 2′-5′-oligoadenylate synthetase/RNase L system for treatment response in chronic hepatitis c. *Journal of Hepatology*, 50(1):9, 2009.

[72] G. Moerkotte and T. Neumann. Dynamic programming strikes back. In J. T.-L. Wang, ed., *Proceedings of the ACM SIGMOD International Conference on Management of Data (SIGMOD 2008)*, Vancouver, Canada, June 2008, pp. 539–552. ACM.

[73] F. Neumann, D. Sudholt and C. Witt. Analysis of different MMAS ACO algorithms on unimodal functions and plateaus. *Swarm Intelligence*, 3(1):35–68, 2009.

[74] T. Neumann. Query simplification: Graceful degradation for join-order optimization. In *Proceedings of the 2009 ACM SIGMOD International Conference on Management of Data (SIGMOD 2009)*, Providence, USA, June 2009. ACM.

[75] T. Neumann and G. Weikum. RDF-3X: a RISC-style engine for RDF. *Proceedings of the VLDB Endowment*, 1(1):647–659, 2008.

[76] T. Neumann and G. Weikum. Scalable join processing on very large RDF graphs. In *Proceedings of the 2009 ACM SIGMOD International Conference on Management of Data (SIGMOD 2009)*, Providence, USA, 2009. ACM.

[77] M. Okabe, K. Anjyo, T. Igarashi and H.-P. Seidel. Animating pictures of fluid using video examples. *Computer Graphics Forum (Proc. EUROGRAPHICS)*, 28(2):677–686, 2009.

[78] K. Panagiotou and A. Steger. Maximal biconnected subgraphs of random planar graphs. In *Proceedings of the 20th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA 09)*, January 4-6, 2009, 2009, pp. 432–440. ACM, SIAM.

[79] K. Patil, P. S. Shelokar, V. K. Jayaraman and B. D. Kulkarni. Ant colony optimization in the direct ordering gene expression data. *Journal of Hybrid Computing Research*, 1(1):10–10, June 2008.

[80] E. Pyrga and S. Ray. New existence proofs for epsilon-nets. In *Proceedings of the twenty-fourth annual symposium on Computational geometry*, College Park, MD, USA, 2008, pp. 199–207. ACM.

[81] T. Ritschel, M. Ihrke, J. R. Frisvad, J. Coppens, K. Myszkowski and H.-P. Seidel. Temporal glare: Real-time dynamic simulation of the scattering in the human eye. *Computer Graphics Forum (Proc. of EUROGRAPHICS)*, 28(3):183–192, 2009.

[82] T. Ritschel, M. Okabe, T. Thormählen and H.-P. Seidel. Interactive reflection editing. *ACM Trans. Graph. (Proc. SIGGRAPH Asia 2009)*, 28(5):129:1–129:7, 2009.

[83] T. Ritschel, K. Smith, M. Ihrke, T. Grosch, K. Myszkowski and H.-P. Seidel. 3d unsharp masking for scene coherent enhancement. In G. Turk, ed., *Proceedings of ACM SIGGRAPH 2008*, Los Angeles, USA, 2008, ACM Transactions on Graphics, vol. 27(3), pp. 1–8. ACM.

[84] M. Rosen-Zvi, A. Altmann, M. Prosperi, E. Aharoni, H. Neuvirth, A. Sönnerborg, E. Schülter, D. Struck, Y. Peres, F. Incardona, R. Kaiser, M. Zazzi and T. Lengauer. Selecting anti-HIV therapies based on a variety of genomic and clinical factors. In *Proceedings of the 16th Annual International Conference on Intelligent Systems for Molecular Biology (ISMB 2008)*,

Toronto, Canada, July 2008, Bioinformatics, vol. 24, pp. i399–406. Oxford University.

[85] O. Sander, F. S. Domingues, H. Zhu, T. Lengauer and I. Sommer. Structural descriptors of protein-protein binding sites. In A. Brazma, S. Miyano and T. Akutsu, eds., *Proceedings of 6th Asia-Pacific Bioinformatics Conference*, Kyoto, Japan, 2008, pp. 79–88. Imperial College Press.

[86] S.-E. Schelhorn, T. Lengauer and M. Albrecht. An integrative approach for predicting interactions of protein regions. *Bioinformatics*, 24(16):i35–i41, 2008.

[87] A. Schlicker and M. Albrecht. FunSimMat: a comprehensive functional similarity database. *Nucleic Acids Research*, 36(Database Issue):D434–D439, 2008.

[88] C. Schoen, J. Blom, H. Claus, A. Schramm-Glück, P. Brandt, T. Müller, A. Goesmann, B. Joseph, S. Konietzny, O. Kurzai, C. Schmitt, T. Friedrich, B. Linke, U. Vogel and M. Frosch. Whole-genome comparison of disease and carriage strains provides insights into virulence evolution in Neisseria meningitidis. *Proceedings of the National Academy of Sciences of the United States of America*, 105(9):3473–3478, 2008.

[89] P. Schüffler, T. Mikeska, A. Waha, T. Lengauer and C. Bock. MethMarker: user-friendly design and optimization of gene-specific dna methylation assays. *Genome Biology*, 10(10):R105, Oct. 2009. PMID: 19804638.

[90] T. Schultz, N. Sauber, A. Anwander, H. Theisel and H.-P. Seidel. Virtual Klingler dissection: Putting fibers into context. *Computer Graphics Forum (Proc. EuroVis)*, 27(3):1063–1070, 2008.

[91] T. Schultz and H.-P. Seidel. Estimating crossing fibers: A tensor decomposition approach. *IEEE Transactions on Visualization and Computer Graphics (Proc. IEEE Visualization)*, 14(6):1635–1642, 2008.

[92] T. Schultz, H. Theisel and H.-P. Seidel. Crease surfaces: From theory to extraction and application to diffusion tensor MRI. *IEEE Transactions on Visualization and Computer Graphics*, 2009. RapidPost, accepted 16 April 2009. doi 10.1109/TVCG.2009.44.

[93] M. Schwarz and M. Stamminger. On predicting visual popping in dynamic scenes. In *Proceedings of Symposium on Applied Perception in Graphics and Visualization 2009*, 2009, pp. 93–100.

[94] M. Seeger. Bayesian inference and optimal design for the sparse linear model. *Journal of Machine Learning Research*, 9:759–813, 2008.

[95] M. Seeger, H. Nickisch, R. Pohmann and B. Schölkopf. Optimization of k-space trajectories for compressed sensing by Bayesian experimental design. *Magnetic Resonance in Medicine*, 2009. In print.

[96] V. Sharma. Complexity of real root isolation using continued fractions. *Theor. Comput. Sci.*, 409(2):292–310, 2008.

[97] V. Sofronie-Stokkermans. Efficient hierarchical reasoning about functions over numerical domains. In K. Berns and T. Breuel, eds., *KI 2008: Advances in Artificial Intelligence*, Kaiserslautern, Germany, 2008, LNAI 5243, pp. 135–143. Springer.

[98] V. Sofronie-Stokkermans. Interpolation in local theory extensions. *Logical Methods in Computer Science*, 4(4):31 pages, October 2008. Special issue of LMCS dedicated to IJCAR 2006.

[99] V. Sofronie-Stokkermans. Locality results for certain extensions of theories with bridging functions. In R. Schmidt, ed., *22nd International Conference on Automated Deduction (CADE-22)*, Montreal, Canada, 2009, Lecture Notes in Artificial Intelligence. Springer. To appear.

[100] F. Suchanek, G. Kasneci and G. Weikum. YAGO: A core of semantic knowledge - unifying WordNet and Wikipedia. In C. L. Williamson, M. E. Zurko and P. J. Patel-Schneider, Peter F. Shenoy, eds., *16th International World Wide Web Conference (WWW 2007)*, Banff, Canada, 2007, pp. 697–706. ACM.

[101] F. Suchanek, G. Kasneci and G. Weikum. YAGO - a large ontology from Wikipedia and WordNet. *Elsevier Journal of Web Semantics*, 6(3):203–217, September 2008.

[102] F. Suchanek, M. Sozio and G. Weikum. SOFIE: A self-organizing framework for information extraction. In *Proceedings of the 18th World Wide Web Conference (WWW 2009)*, Madrid, Spain, 2009. ACM.

[103] A. Tevs, M. Bokeloh, M. Wand, A. Schilling and H.-P. Seidel. Isometric registration of ambiguous and partial data. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009)*, Miami Beach, Florida, USA, 2009. IEEE Computer Society.

[104] M. Theobald, M. AbuJarour and R. Schenkel. TopX 2.0 at the INEX 2008 Efficiency Track. In S. Geva, J. Kamps and A. Trotman, eds., *INEX, 2008, LNCS 5631*, pp. 224–236. Springer.

[105] M. Theobald, H. Bast, D. Majumdar, R. Schenkel and G. Weikum. TopX: Efficient and versatile top-k query processing for semistructured data. *The VLDB Journal*, 17(2):81–115, January 2008.

[106] M. Theobald, R. Schenkel and G. Weikum. The TopX DB&IR engine (demo). In N. Koudas, ed., *2007 ACM SIGMOD International Conference on Management of Data, Beijing*, 2007, pp. 1141–1143. ACM.

[107] T. Thormählen and H.-P. Seidel. 3d-modeling by ortho-image generation from image sequences. *ACM Transactions on Graphics (Proc. ACM SIGGRAPH)*, 27(3):86:1–86:5, August 2008.

[108] H. R. Tiwary and K. Elbassioni. On the complexity of checking self-duality of polytopes and its relations to vertex enumeration and graph isomorphism. In *Symposium on Computational Geometry 2008*, College Park, MD, USA,, 2008, pp. 192–198. ACM.

[109] M. Trignetti, T. Sing, V. Svicher, M. Santoro, R. Forbici, R. D'arrigo, M. Bellocchi, M. Santoro, P. Marconi, M. Zaccarelli, M. Tortta, R. Bellagamba, P. Narciso, A. Antinori, C. Perno, T. Lengauer and F. Ceccherini-Silberstein. Dynamics of NRTI resistance mutations during therapy interruption. *AIDS Research and Human Retroviruses*, 25(1):7, 2009.

[110] G. Van Ooijen, G. Mayr, M. Albrecht, B. J. C. Cornelissen and F. L. W. Takken. Transcomplementation, but not physical association of the CC-NB-ARC and LRR domains of tomato R protein Mi-1.2 is altered by mutations - in the ARC2 subdomain. *Molecular Plant*, 1(3):401–410, 2008.

[111] G. Van Ooijen, G. Mayr, M. M. A. Kasiem, M. Albrecht, B. J. C. Cornelissen and F. L. W. Takken. Structure-function analysis of the NB-ARC domain of plant disease resistance proteins. *Journal of Experimental Botany*, 59(6):1383–1397, 2008.

[112] M. Wand, B. Adams, M. Ovsjanikov, A. Berner, M. Bokeloh, P. Jenke, L. Guibas, H.-P. Seidel and A. Schilling. Efficient reconstruction of non-rigid shape and motion from real-time 3d scanner data. *ACM Transactions on Graphics*, 2009.

[113] C. Weidenbach, D. Dimova, A. Fietzke, M. Suda and P. Wischnewski. Spass version 3.5. In R. A. Schmidt, ed., *22nd International Conference on Automated Deduction (CADE-22)*, Montreal, Canada, August 2009, *LNAI 5663*, pp. 140–145. Springer.

[114] N. Weinhold, O. Sander, F. S. Domingues, T. Lengauer and I. Sommer. Local function conservation in sequence and structure space. *PLoS Computational Biology*, 4:e1000105, July 2008.

[115] C. Welsch, F. S. Domingues, S. Susser, I. Antes, C. Hartmann, G. Mayr, A. Schlicker, C. Sarrazin, M. Albrecht, S. Zeuzem and T. Lengauer. Molecular basis of telaprevir resistance due to V36 and T54 mutations in the NS3-4A protease of HCV. *Genome Biology*, 9(1):R16, 2008.

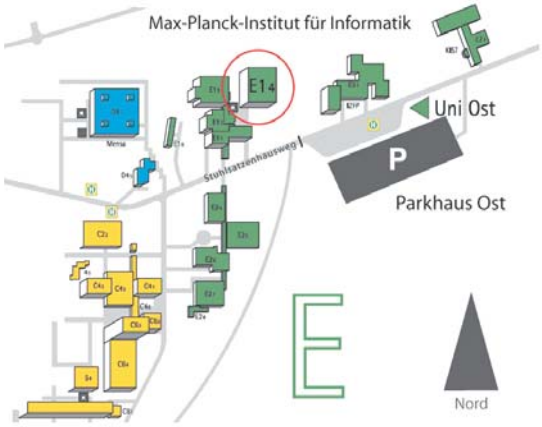
[116] C. Xu, A. Karrenbauer, K. M. Soh and C. Codrea. Consecutive multiline addressing: A scheme for addressing pmoleds. *Journal of the Society for Information Display*, 16(2):211–219, 2008.

[117] H. Zhu, I. Sommer, T. Lengauer and F. S. Domingues. Alignment of non-covalent interactions at protein-protein interfaces. *PLoS ONE*, 3(4):e1926, April 2008.

Wege zum Institut

Das Max-Planck-Institut für Informatik (Gebäude E1 4) befindet sich auf dem Campus der Universität des Saarlandes etwa 5 km nordöstlich vom Zentrum der Stadt Saarbrücken im Wald nahe bei Dudweiler.

Saarbrücken besitzt einen eigenen Flughafen (Saarbrücken-Ensheim) und ist mit Auto-, Shuttle-Bus- und Zugverbindungen an die Flughäfen Frankfurt und Luxemburg angebunden. Zugstrecken verbinden Saarbrücken im Stundentakt innerhalb Deutschlands. Regelmäßig verkehrende Züge schaffen eine Anbindung an die Städte Metz, Nancy und Paris. Autobahnen führen nach Mannheim/Frankfurt, Luxembourg/Trier/Köln, Strasbourg und Metz/Nancy/Paris.



Sie erreichen den Campus...

von Saarbrücken-Ensheim, Flughafen

mit dem Taxi in ungefähr 20 Minuten

von Saarbrücken, Hauptbahnhof

mit dem Taxi in ungefähr 15 Minuten
mit dem Bus in ungefähr 20 Minuten
Richtung *„Dudweiler-Dudoplatz“* oder *„Universität Campus“*
Ausstieg *„Universität Mensa“*
alternativ Ausstieg *„Universität Campus“*

von Frankfurt oder Mannheim über die Autobahn A6

Abfahrt *„St.Ingbert-West“*
den weißen Schildern *„Universität“* zum Campus folgend

von Paris über die Autobahn A4

Abfahrt *„St.Ingbert-West“*
den weißen Schildern *„Universität“* zum Campus folgend

IMPRESSUM

Herausgeber
Max-Planck-Institut für Informatik
Campus E1 4
D-66123 Saarbrücken

Redaktion & Koordination
Alice McHardy
Manuel Lamotte-Schubert
Thomas Lengauer
Kurt Mehlhorn
Jennifer Müller
Frank Neumann
Michael Schwarz
Hans-Peter Seidel
Martin Theobald
Uwe Waldmann
Christoph Weidenbach
Gerhard Weikum
Elena Zotenko

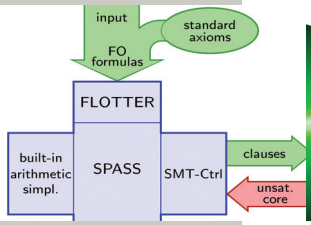
Kontakt
Max-Planck-Institut für Informatik
Telefon +49 681 9325-0
Telefax +49 681 9325-719
Email info@mpi-inf.mpg.de
Internet http://www.mpi-inf.mpg.de

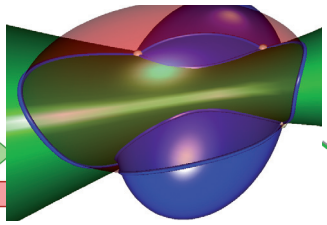
Berichtszeitraum
1. Januar 2007 bis 31. Dezember 2008

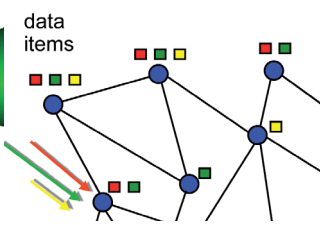
Gestaltung
Behr Design | Saarbrücken

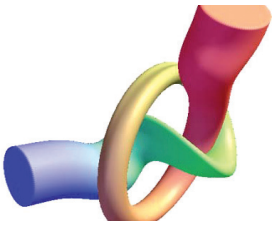
Druck
Bliesdruckerei | Blieskastel

max planck institut
informatik









Max-Planck-Institut für Informatik
Campus E1 4
D-66123 Saarbrücken

Telefon +49 681 9325-0
Telefax +49 681 9325-719
Email info@mpi-inf.mpg.de
Internet http://www.mpi-inf.mpg.de

max planck institut
informatik