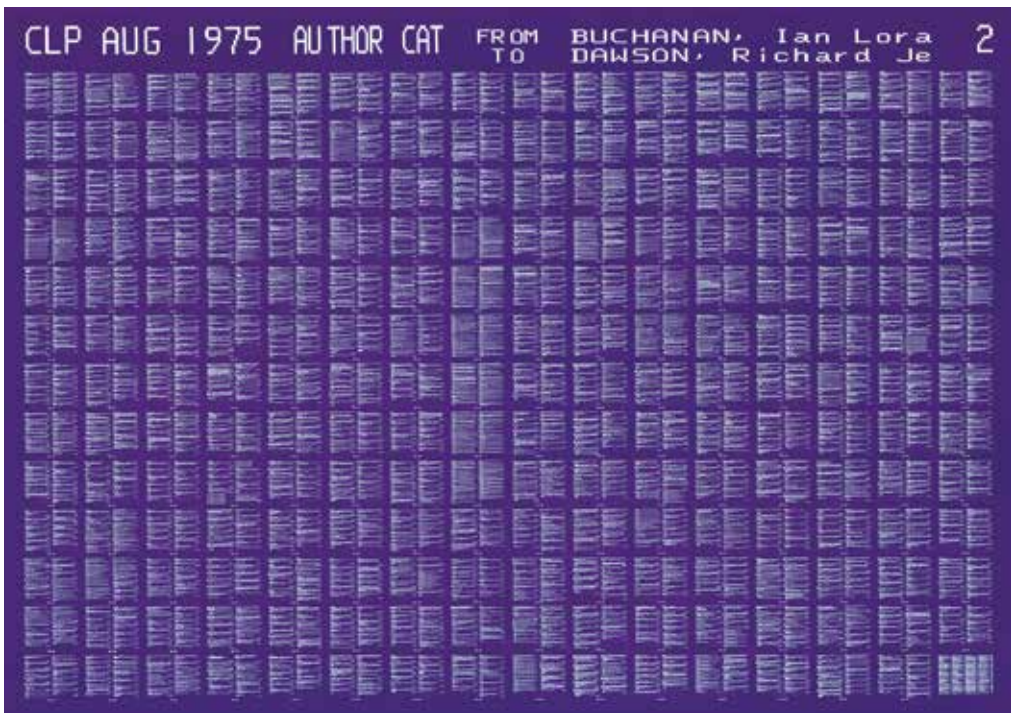
Wie bewältigen wir die Datenflut?

Neues DFG-Schwerpunkt-Programm zu Algorithmen für Big Data

Wir sammeln und produzieren jedes Jahr eine exponentiell wachsende Zahl von Daten: Twitter-User generieren täglich über 300 Millionen Tweets und eine vergleichbar große Zahl von Bildern wird täglich von Google-Nutzern hochgeladen. Wissenschaftliche Experimente wie der Large Hadron Collider bei Genf produzieren jährlich rund 15 Petabytes, eine unvorstellbar große Zahl mit 15 Nullen. In vielen Gebieten nimmt die Datenflut aufgrund sinkender Kosten rasant zu – beispielsweise ist die Masse der Daten aus DNA-Sequenzierungen schneller angewachsen als die Entwicklung der Hard- und Software zu ihrer Verarbeitung. „Wir leben in einer Big Data World, in der das wirkliche Problem nicht mehr das Sammeln der Daten ist, sondern die Bewältigung ihrer ungeheuren Masse“, fasst der Informatiker Prof. Ulrich Meyer die Situation zusammen. Er ist Professor für Algorithm Engineering und wissenschaftlicher Koordinator eines neuen Schwerpunktprogramms der Deutschen Forschungsgemeinschaft, das sich zum Ziel gesetzt hat, Algorithmen für die Verarbeitung großer Datenmengen zu entwickeln.



Das schnelle Altern von Speichermedien: Mikrofiche aus dem Jahre 1975.
Foto: ullstein bild – NMSI/Science Museum / Science Museum

Hardware hält nicht Schritt mit Datenmenge

Eines der Probleme besteht darin, dass die Datenflut schneller anwächst als die Zahl der elementaren Schaltkreiselemente auf einem Computerchip. Diese verdoppelt sich dem Moore'schen Gesetz entsprechend alle 18 Monate. Die Datenmenge verdoppelt sich dagegen je nach Datentyp deutlich schneller, wobei viele Daten auch repliziert werden. Aber selbst dann, wenn die Entwicklung der Hardware mit dem Datenwachstum Schritt halten könnte, bedeutete dies nicht, dass die Daten gleichbleibend schnell verarbeitet werden könnten. „Dass sich die Geschwindigkeit der Rechenoperationen im gleichen Maße erhöhte, wie die Anzahl der elementaren Schaltkreiselemente zunahm, gehört der Vergangenheit an“, erklärt Meyer: „Heute braucht man nicht nur mehr Speicherplatz, sondern auch mehrere parallel arbeitende Prozessoren, um die Daten in einer vertretbaren Zeit zu verarbeiten.“

Und dazu benötigt man neue Algorithmen, die nicht nur eine parallele Datenverarbeitung leisten, sondern auch bestenfalls linear mit dem Dateninput skalieren. Das heißt: Wenn sich die Zahl der Daten verdoppelt, sollte sich die Anzahl der Rechenoperationen idealerweise auch nur verdoppeln – und diese dann auch noch gut auf parallele Prozessoren verteilen lassen. Das ist längst nicht bei allen Problemen der Fall. Meyer nennt als Beispiel Algorithmen, die für den Navi die kürzeste Strecke zwischen zwei Orten berechnen. Die Orte und die dazwischen liegenden Strecken werden in der Informatik als Graphen repräsentiert – ein Netzwerk aus Knotenpunkten, die mit Strichen verbunden sind.

Anwendungsfall Routenplanung

Die einfachste Version des Problems hat zwar eine fast lineare Laufzeit, aber auch nur falls der Graph in den Hauptspeicher passt. Gleichzeitig fehlen wirklich effiziente Parallelisierungen. Für kompliziertere Varianten, wie zum Beispiel das Finden der schnellsten Verbindung bei Treibstoffbeschränkungen oder der Kombination verschiedener Transportmittel, kann der Rechenaufwand schlimmstenfalls sogar exponentiell mit der Zahl der Knoten anwachsen. Auf diesem Gebiet sind in den vergangenen Jahren bereits Fortschritte erzielt worden. 2004 war ein Straßennetz von West-Europa, das rund 20 Millionen Knoten enthielt, noch eine Herausforderung: In einem solchen Graphen die kürzeste Strecke zu finden, dauerte auf einem Webserver zu lang und war zu teuer. Heute sind die Berechnungen dank verbesserter Algorithmen um sechs Größenordnungen schneller geworden.

„Wir sind jedoch weit davon entfernt, das Problem in allen Ausprägungen gelöst zu haben“, gibt Meyer zu bedenken. Inzwischen arbeiten Informatiker mit Graphen, die mehrere Billionen Knoten haben. Sie wollen zusätzlich den Verkehrsfluss in die Routenplanung einbeziehen und mit einer wachsenden Anzahl mobiler GPS-Nutzer kommunizieren, die laufend ihre Position senden. Die öffentlichen Verkehrsmittel sollen in die Routenplanung einbezogen werden. Um Wartezeiten zu minimieren, müssen die Programmierer Fahrpläne berücksichtigen. Und schließlich wollen sie auch auf die Vorlieben des Nutzers eingehen.

DFG-Förderung über sechs Jahre

Zeit und Kosten sparende Algorithmen für Graphen zu entwickeln ist nur eines von vielen praktischen Problemen, die Ulrich Meyer und seine Kollegen von insgesamt acht deutschen Universitäten in den nächsten sechs Jahren in Angriff nehmen wollen. Die Deutsche Forschungsgemeinschaft fördert das Vorhaben mit insgesamt 4,9 Millionen Euro, wobei 878.000 Euro nach Frankfurt gehen. 596.000 Euro erhält Ulrich Meyer für die Koordinierung des Schwerpunktprogramms und projektübergreifende Aktionen, 282.000 Euro für das eigene Forschungsprojekt „Big-Data-DynAmO“. In diesem Projekt geht es unter anderem darum, Algorithmen für riesige Graphen zu entwickeln, die sich dynamisch ändern. Bei der Routenplanung wäre dies beispielsweise die Berücksichtigung von Staus, die eine schnelle Routenänderung erforderlich machen. Hier soll der Ansatz verfolgt werden, auf bereits berechneten Lösungen aufzubauen, anstatt die Berechnung komplett von vorn zu beginnen. „In manchen Big Data Szenarien müssen wir uns aber von einer exakten Lösung verabschieden und stattdessen zwischen dem Rechenaufwand und der Qualität einer genäherten Lösung abwägen“, erläutert Meyer das allgemeine Vorgehen.

Anfang Juni trafen sich mehrere Projektleiter und Mitarbeiter zum Projektstart auf dem Campus Riedberg. Ein wichtiges Anliegen ist ihnen die multidisziplinäre Herangehensweise an Probleme aus verschiedensten Praxisbereichen von der Genomforschung über Suchmaschinen und Routenplanern bis hin zu Kommunikationsnetzwerken wie Facebook. „Diese Communities haben oft ähnliche Probleme bei der Bewältigung großer Datenmengen, kooperieren aber bisher nur wenig miteinander“, so Meyer. In dem DFG-Schwerpunktprogramm soll wissenschaftliche Grundlagenforschung eng mit den Problemen der Anwender verzahnt werden. Wichtig ist den Projektleitern dabei auch die Ausbildung der nächsten Generation von Informatikern und explizit auch Informatikerinnen, die der Datenflut künftig Herr werden müssen. *Anne Hardy*

Infos zu den Einzelprojekten unter:
➤ www.big-data-spp.de

Wider disziplinäre Trennung

Interdisziplinäre Tagung zeigt Potential der Netzwerkforschung für Praxis auf

In einer Tagung im Schader-Forum Darmstadt am 12. und 13. Juni diskutierten mehr als 120 Wissenschaftler und Praktiker unterschiedlicher Bereiche und Disziplinen die Möglichkeiten netzwerkanalytischer Perspektiven. Die Netzwerkforschung selbst kann als Ensemble von, im Vergleich zu traditionellen Methoden, hochpotenten Denk- und Analyseverfahren bezeichnet werden. Die Inhalte sind vielfältig: Netzwerkforschung hilft beispielsweise aufzuklären, wie sich Innovationen in Unternehmen etablieren und warum in der Wirtschaft so eng zusammengearbeitet wird, wie die sozialen Grundlagen für die Entstehung und Diffusion von Technologie und deren Anwendung beschaffen sind oder warum die Stakeholderkonstellation kaum „vernünftige“ nachhaltige Wirtschaft zulässt.

Die 40 Vorträge hielten vor allem Wissenschaftler aus unterschiedlichen Disziplinen wie Geographie, Soziologie, Wirtschaftswissenschaft, Politologie und Informatik. Dabei wurde deutlich, dass die Gedankenwelt der vorwiegend aus der Soziologie stammenden Netzwerkforschung die Grenzen dieser Disziplinen durchlöchert hat. Die Grenzen lassen sich vor allem in der institutionellen Einbindung und den unterschiedlichen Karrierewegen ausmachen, viel seltener aber an den Forschungsthemen und den methodischen Zugängen.

Eine Podiumsdiskussion mit Praxisvertretern zeigte die Bedeutung der Forschung für die Praxis auf. Teilnehmer waren Franz Grubauer, Oberkirchenrat und zuständig für Statistik bei der Ev. Kirche Hessen-Nassau, Stefan Klingelhöfer, Personalchef der Lufthansa City Line, und Rüdiger Feibel, Geschäftsführer eines Zusammenschlusses von zahlreichen mittelständischen Zahntechnikunternehmen. In der Diskussion wurde der Wert der Netzwerkforschung für die Praxis aufgezeigt. So hat die ev. Kirche eine große Gemeindestudie in Auftrag gegeben, die zu verstehen hilft, wer mit wem über religiöse Inhalte spricht. In der City Line spielt die Netzwerkforschung u.a. bei der Nachfolgeplanung für Managerposten eine wichtige Rolle.

Die Tagung entsprang einer Zusammenarbeit der Soziologie der Goethe-Universität (Christian Stegbauer), der Wirtschaftsgeographie der Karls-Universität Heidelberg (Johannes Glückler) und der Schader-Stiftung in Darmstadt.

Christian Stegbauer