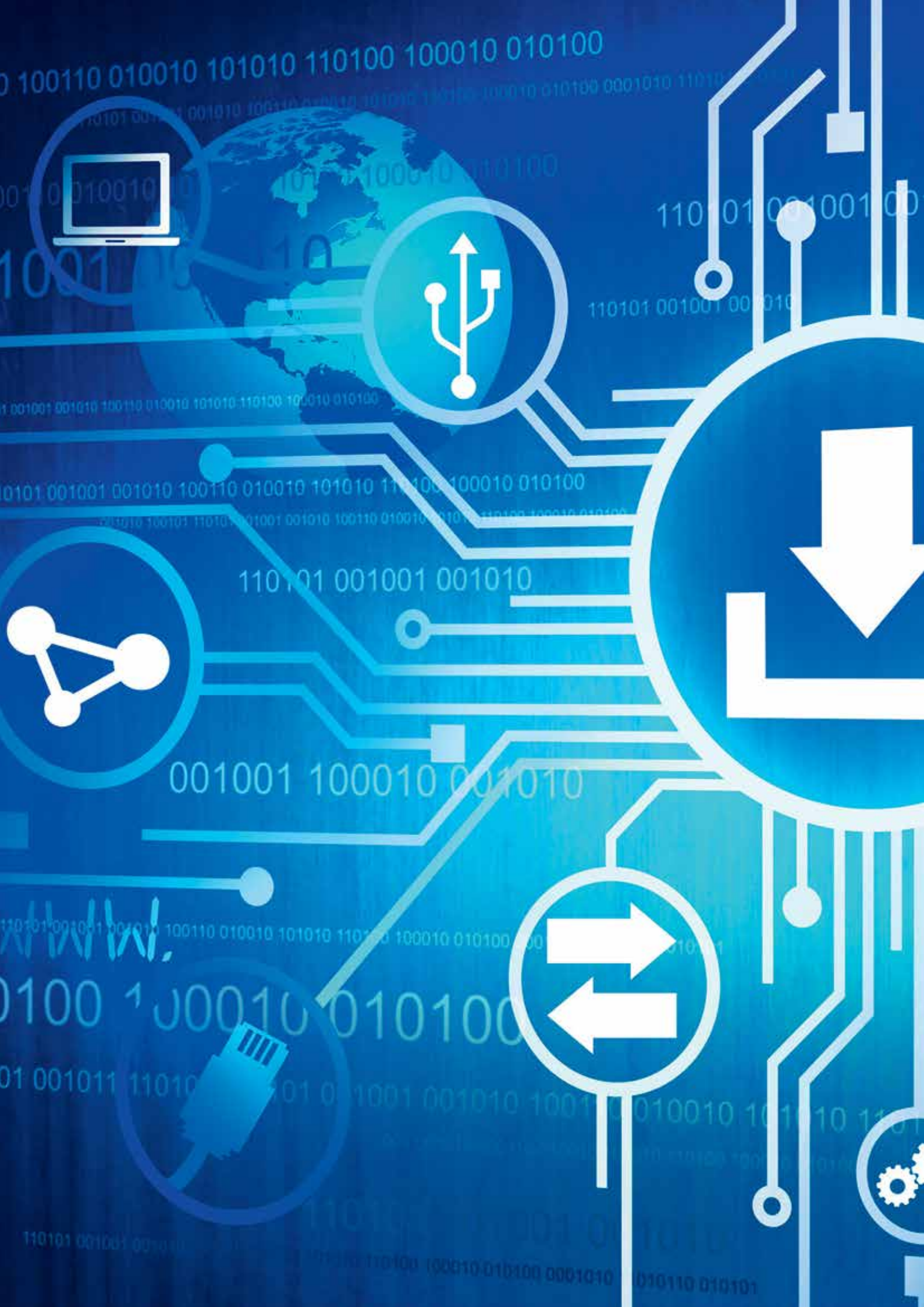

Algorithmische Überlebensstrategien in der Datenflut

DFG-Schwerpunkt »Algorithms for Big Data« bündelt Expertisen bundesweit

von Ulrich Meyer

In der Big-Data-Welt ist nicht mehr die Akquise, sondern die Verarbeitung von Daten das größte Problem. Prof. Dr. Ulrich Meyer, Koordinator des DFG-Schwerpunktprogramms »Algorithms for Big Data«, erläutert einige Herausforderungen beim Umgang mit großen Datenmengen.

Computersysteme durchdringen alle Bereiche menschlicher Aktivität. Immer schneller erheben, verarbeiten und versenden sie riesige Datenmengen untereinander. Als Konsequenz leben wir in einer Big-Data-Welt, in der das Informationsvolumen exponentiell zunimmt und die eigentlichen Probleme nicht mehr in der Akquise hinreichend vieler Daten, sondern eher in der Handhabung ihrer Fülle und ihres ungestümen Wachstums liegen. Da der Geschwindigkeitszuwachs einzelner Prozessorkerne im Wesentlichen zum Stillstand gekommen ist, setzt die Hardwareindustrie auf immer mehr Berechnungskerne pro Board oder Grafikkarte und investiert in neue Speichertechnologien. Das bedeutet, dass unsere Algorithmen, also die von Menschen erdachten Berechnungsvorschriften, nach denen Computer konkrete Probleme lösen, massiv parallel werden und auf Datenlokalität setzen müssen.

Dass es mit der Parallelität zuweilen trickreich ist, weiß schon der Volksmund: Machen sich zu viele Köche gleichzeitig an einem Brei zu schaffen, haben sie die Tendenz, selbigen zu verderben. Glücklicherweise ist das bei algorithmischen Fragestellungen nicht immer der Fall; das Sortieren großer Datenmengen ist zum Beispiel sehr gut parallelisierbar. Dennoch kann man für realistische Prozessorzahlen die tatsächliche Sortierzeit nicht beliebig verringern. Ganz ähnlich wie man mit vielen Herden und Schüsseln zwar mehr Brei pro Stunde kochen kann als mit nur einem Herd, aber die Zeit für die erste fertige Portion auch bei vielen Herden nicht beliebig kurz wird. Für manche Probleme kann man sogar mathematisch beweisen, dass es kaum etwas bringt, wenn zu viele Prozessoren gleichzeitig in der Datensuppe rühren.

Von Festplatten und Vorratskammern

Eine weitere Herausforderung, mit der sich auch die Arbeitsgruppe des Autors beschäftigt, besteht darin, durch neue Algorithmen und Datenstrukturen eine bessere Datenlokalität zu erzielen. Eine solche liegt vor, wenn sich die Speicherzugriffe der Algorithmen auf wenige Stellen des Speichers konzentrieren und/oder im zeitlichen Verlauf auf Nachbarzellen von zuvor besuchten Speicherzellen zugegriffen wird. Datenlokalität macht sich vor allem bezahlt, wenn die Daten aufgrund ihrer Größe nur noch auf langsamen externen Speichermedien wie Festplatten gehalten werden können, bei denen der Zugriff auf einen ganzen Datenblock kaum langsamer ist als das Lesen einer einzelnen Speicherzelle. Je größer der Hauptspeicher, umso mehr Datenblöcke können somit für zukünftig schnellen Zugriff zwischengespeichert werden.

Eine einfache Analogie aus dem täglichen Leben ergibt sich beim Getränkekauf: Statt nach jeder ausgetrunkenen Flasche des Lieblingsgetränks eine neue Einzelflasche aus dem entfernten Getränkemarkt zu beschaffen, wird

Jahrestreffen 2014 des Big-Data-Schwerpunkts in Frankfurt

gleich eine ganze Kiste heimgefahren und in der schnell zugänglichen Vorratskammer deponiert. Da die Kammer aber eher klein ist, werden selten benötigte Güter typischerweise nicht in größeren Mengen vorgehalten.

Kompliziert wird es, wenn sich das Verbrauchsverhalten laufend und unvorhersehbar ändert oder die Kammer in einer Wohngemeinschaft für mehrere Nutzer mit unterschiedlichen Präferenzen gleichzeitig zur Verfügung stehen soll. Hier braucht es dann deutlich mehr Abstimmung und Organisation. Ganz ähnlich verhält es sich mit auf Speicherzugriffe optimierten Algorithmen. Nur dass hier außer der Organisation des Hauptspeichers auch die externen Datenstrukturen und der gesamte Programmablauf geeignet angepasst werden. Das ist in etwa so, als ob der Getränkekunde nun auch die Logistik und Warenplatzierung seiner Einkaufsmärkte für seine Zwecke optimieren könnte.

Immer öfter ist es im Kontext großer Datenmengen gar nicht mehr möglich, tatsächlich alle verfügbaren Daten zu betrachten. Stattdessen kann zum Beispiel eine zufällige Auswahl getroffen werden. Wie folgenreich diese »Mut-zur-Lücke«-Strategie ist, hängt dann sowohl von dem eigentlichen Problem als auch von der konkreten (oft adaptiven) Auswahlstrategie ab. Beispielsweise mag es zur Analyse der Struktur sozialer Netzwerke reichen, sich auf eine zufällige Teilmenge der Nutzer zu beschränken. Bei der Hochwasservorhersage basierend auf einer (zu) groben Auswahl der Milliarden von Bodenhöhenmesspunkten könnte jedoch zum Beispiel ein Deich (oder Durchbrüche desselben) durch das Raster fallen und somit zu vollkommen falschen Prognosen führen.

Um Herausforderungen wie den oben beschriebenen erfolgreich zu begegnen, braucht es typischerweise neue algorithmische Ideen. Es geht hier also gar nicht unbedingt und ausschließlich um die so oft beschworenen neuen Möglichkeiten, den vorhandenen Datenreichtum zu nutzen, sondern schlicht um algorithmische Überlebensstrategien in der Datenflut.

In diesem Sinne hat eine Initiatorengruppe bestehend aus acht Professoren an deutschen Universitäten und Max-Planck-Instituten unter Federführung des Autors bei der Deutschen Forschungsgemeinschaft (DFG) das Schwer-

AUF DEN PUNKT GEBRACHT

- Seit 2014 fördert die Deutsche Forschungsgemeinschaft (DFG) das auf sechs Jahre angelegte Schwerpunktprogramm »Algorithms for Big Data« unter Federführung der Goethe-Universität. Es bündelt die Expertise von 14 deutschen Forschungseinrichtungen aus unterschiedlichsten Bereichen.
- Ziel ist es, effiziente Algorithmen und Datenstrukturen zur Bewältigung großer Datenmengen zu entwickeln.
- Das Durchlaufen des »Algorithm Engineering«-Zyklus schlägt eine Brücke zwischen Theorie und Praxis.

SPP BIG DATA

punktprogramm »Algorithms for Big Data« beantragt, das 2013 mit einem Gesamtfördervolumen von 4,9 Millionen Euro für die erste Dreijahresphase eingerichtet wurde. Der überwiegende Teil dieser Mittel fließt in die Doktorandenausbildung an den einzelnen Projektstandorten sowie deren Vernetzung.

Ein Baukasten verbesserter Algorithmen für die Praxis

Das Schwerpunktprogramm soll die Expertise aus verschiedenen Gebieten bündeln. Einerseits müssen aktuelle Hardwareentwicklungen und technologische Herausforderungen adäquat in bessere Berechnungsmodelle einfließen. Andererseits sollen sowohl allgemeine als auch anwendungsspezifische algorithmische Probleme, die sich aus der Größe der Daten ergeben, identifiziert und klassifiziert werden. Vor diesem Hintergrund ist geplant, einen Baukasten verbesserter Algorithmen und Datenstrukturen für große Datenmengen zu entwickeln, bei dem es nicht nur um theoretische Resultate geht, sondern der volle »Algorithm Engineering«-Zyklus durchlaufen werden soll. Algorithm Engineering umfasst den Entwurf, die theoretische Analyse, Implementierung und experimentelle Auswertung von Algorithmen und soll die Lücke schließen zwischen hocheffizienten, aber schwer zugänglichen theoretischen Ansätzen und von Praktikern benutzten einfachen, aber oft ineffizienten Lösungen.

Konkrete algorithmische Big-Data-Herausforderungen beinhalten das Ausnutzen von Parallelität (Multicores, GPUs, Clouds) und Speicherhierarchien (Festplatten, Flashspeicher, Caches), den Umgang mit kontinuierlichen massiven Datenaktualisierungen, die Verarbeitung komprimierter und verschlüsselter Daten, die Approximation und Online-Verarbeitung bei beschränkten Ressourcen oder die Reduktion des Energieverbrauchs durch algorithmische Maßnahmen. Was die Initiative von den meisten früheren Arbeiten unterscheidet, ist der Ansatz, nicht nur bestimmte isolierte Probleme anzupacken, sondern Verbundlösungen für gleich mehrere Aspekte zu suchen.

Beispielsweise soll für Textindizierungsprobleme untersucht werden, wie durch die gemeinsame Ausnutzung von Parallelität, Speicherhierarchien, Besonderheiten der Daten und neuer algorithmischer Techniken eine bessere Gesamtperformanz erreicht werden kann. Von Anfang an wird durch Kooperationen mit Anwendungsfeldern (zum Beispiel aus der Biologie oder den Informationswissenschaften) sichergestellt, dass neben theoretischer Grundlagenforschung auch anwendungsrelevante Fragen zum Nutzen mehrerer Communities bearbeitet werden.

Aus den rund 40 eingegangenen Projektanträgen hat eine internationale Gutachterkommission Angang 2014 fünfzehn Projekte für die erste Förderperiode des Schwerpunktprogamms ausgewählt. Ein weiteres, thematisch passendes Projekt aus der DFG-Einzelförderung wurde dem Schwerpunkt später assoziiert. Die nunmehr 16 wissenschaftlichen Projekte umfassen die Bereiche Technologische Herausforderungen, Netzwerke, Optimierung, Sicherheit, Textanwendungen und Bioanwendungen.

Hier einige Beispiele: Eine Fragestellung aus dem Bereich technologische Herausforderungen betrifft neue algorithmische Methoden, um bei diskret wählbaren Prozessorgeschwindigkeiten adaptiv einen guten Kompromiss aus Geschwindigkeit und Energieverbrauch zu erzielen. Ein repräsentatives Netzwerkproblem ist die Suche nach optimalen Routen in großen Verkehrsnetzwerken, bei denen sich die Reisedauer zwischen den Netzwerknoten je nach Belastung und Zeit ständig ändert. Und im Bereich Sicherheitsaspekte wird beispielsweise der Frage nachgegangen, wie große Datenmengen verschlüsselt auf externen Servern abgelegt werden können, damit der externe Dienstleister auf diesen Daten zwar die Berechnungen des Kunden ausführen, dabei die tatsächlichen Daten aber nicht ausspähen kann. Neben den wissenschaftlichen Fachprojekten existiert auch noch ein Koordinierungsprojekt, das von Frankfurt aus die Zusammenarbeit des Schwerpunkts unterstützt.

Besonderer Wert wird innerhalb des Schwerpunktprogramms auf die Ausbildung von Doktoranden gelegt. Ein erstes gemeinsames Treffen aller Teilnehmer fand dann im Herbst 2014 in Frankfurt statt (siehe Foto Seite 64). Die Doktoranden trafen sich im Mai in Montabaur. In kleineren themenbezogene Workshops und einer Summer School in Frankfurt soll die Vernetzung weiter gestärkt werden. So arbeitet die nachwachsende Forschergeneration bereits daran, die zunehmende Datenflut in unterschiedlichsten Bereichen auch in Zukunft sinnvoll verarbeiten und nutzen zu können. ●

Der Autor

Prof. Dr. Ulrich Meyer, Jahrgang 1971, promovierte 2002 an der Universität des Saarlandes und dem Max-Planck-Institut für Informatik. Aufenthalte als Gastwissenschaftler führten ihn an die Ungarische Akademie der Wissenschaften in Budapest und an die private Duke University in Durham/North Carolina, USA. Von 2005 bis 2007 war er Senior Researcher am MPI für Informatik. 2007 folgte er dem Ruf als Professor für Algorithm Engineering an der Goethe-Universität. Seit Ende 2014 ist er Studiendekan in der Informatik.

umeyer@cs.uni-frankfurt.de

www.big-data-spp.de