DFG-Graduiertenkolleg 1324/2

MODELLBASIERTE ENTWICKLUNG VON TECHNOLOGIEN FÜR SELBSTORGANISIERENDE DEZENTRALE INFORMATIONSSYSTEME IM KATASTROPHENMANAGEMENT (METRIK)

http://www.gk-metrik.de

Sprecher

PROF. DR. SC. NAT. JOACHIM FISCHER
Tel.: (030) 2093 3109
e-mail: fischer@informatik.hu-berlin.de

Sekretariat

GABRIELE GRAICHEN
e-mail: graichen@informatik.hu-berlin.de

Doktoranden

DIPL.-INF. OSWALD BERTHOLD bis 08/2015
DIPL.-INF. HARTMUT LACKNER
M. SC. TOBIAS RAWALD
M. SC. LARS GEORGE
DIPL.-INF. FABIAN FIER
DIPL.-INF. HEINRICH MELLMANN
M.SC. JOHANNES SCHREYER
M.SC. MARTIN SCHMIDT

POSTDOCS

DR. BRUNO CADONNA bis 07/15
DR. PANAGIOTIS BOUROS bis 03/15

Assoziierte

DIPL.-INF. BJÖRN LICHTBLAU
DIPL.-INF. LARS DÖHLING
DIPL.-GEOGR. CARSTEN KRÜGER
DIPL.-INF. MATTHIAS SAX
DIPL.-INF. MICHAEL FREY
DIPL.-ING. VLADICA SARK
DIPL.-PHYS. CHRISTIAN BLUM
M. SC. ENG. JOANNA GEIBIG
DIPL.-GEOGR. ANDREAS REIMER
DIPL.-INF. CHRISTOPH WAGNER
DIPL.-INF. ANDREAS DITTRICH

M. Sc. Jens Nachtigall
Dipl.-Inf. Arif Wider
Dipl.-Inf. Martin Kost
M.Sc. Daniel Cagara

Forschungsstudenten

Münchmeyer, Jannes bis 03/15
Dudek, Robert bis 03/15
Schröder, Sven bis 03/15
Bechtle, Sarah bis 03/15

Betreuende Hochschullehrer

Prof. Dr. Doris Dransch, Geo-Informationsmanagement und -Visualisierung
Prof. Dr. Joachim Fischer, Systemanalyse
Prof. Johann-Christoph Freytag, PhD., Datenbanken und Informationssysteme
Prof. Dr. Verena Hafner, Kognitive Robotik
Prof. Dr. Tobia Lakes, Geomatik
Prof. Dr. Ulf Leser, Wissensmanagement in der Bioinformatik
Prof. Dr. Miroslaw Malek, Rechnerorganisation und Kommunikation, bis 2012
Prof. Dr. Jens-Peter Redlich, Systemarchitektur
Prof. Dr. Alexander Reinefeld, Parallele und verteilte Systeme
Prof. Dr. Wolfgang Reisig, Theorie der Programmierung
Prof. Dr. Holger Schlingloff, Spezifikation, Verifikation und Testtheorie

Assoziierte Hochschullehrer

Prof. Mesut Günes, Verteilte, Eingebettete Systeme
Prof. Dr.-Ing. Eckhard Graß, Drahtlose Breitbandkommunikationssysteme
Prof. Dr. Björn Scheuermann, Technische Informatik

Seit dem Start des interdisziplinären Graduiertenkollegs im Jahre 2006 wurden insgesamt 51 Promotionsprojekte auf den Weg gebracht, die unterschiedlich finanziert worden sind. Sämtliche Projekte waren unabhängig von der Art der Finanzierung sowohl in das Gesamtforschungskonzept integriert als auch dem einheitlichen Qualitätsmanagement von METRIK untergeordnet.

Aus den Mitteln des Graduiertenkollegs konnten den Doktorandinnen und Doktoranden 18 Stipendien und 14 Mitarbeiter-Stellen angeboten werden. Für Postdoktorandinnen und Postdoktoranden standen vier weitere Stellen bereit. Zwei Doktoranden profitierten im Vorfeld der Aufnahme ins Graduiertenkolleg von bewilligten Qualifizierungsstipendien. Darüber hinaus befanden sich in der zweiten Antragshälfte 16 studentische Hilfskräfte in verschiedenen Aufgabenbereichen zur Unterstützung technologischer Problemstellungen im Einsatz. Zusätzlich kamen

zwei weitere Finanzierungsarten zum Einsatz. Zum einen ermöglichte eine Mischfinanzierung der DFG mit ergänzenden parallel laufenden Drittmittelprojekten die Einrichtung weiterer 5 Stipendien- und 7 Mitarbeiterstellen. Zum anderen wurde METRIK um 7 weitere komplett fremd finanzierte Stellen ergänzt und stellt mit den 51 Promotions- und 4 PostDoc-Stellen das bislang umfangreichste Forschungsprojekt am Institut dar.

Im Berichtszeitraum haben 7 METRIK-Doktorandinnen und Doktoranden ihre Dissertationen erfolgreich verteidigt.

1) Mihal Brumbulli, 17. März 2015
   Model-driven development and simulation of distributed communication systems
2) Andreas Dittrich, 20. März 2015,
   Service availability and discovery responsiveness – a user-perceived view on service dependability
3) Reimer, Andreas, 1. Juni 2015,
   Cartographic Modelling for Automated Map Generation
4) Siamak Haschemi, 14. Juli 2015,
   Model-based testing of dynamic component systems
5) Joanna Geibig, Oktober 2015,
   Peer-to-Peer Algorithms in Wireless Ad-Hoc Networks for Disaster Management
6) Arif Wider, 18. November 2015,
   Model transformation languages for domain-specific workbenches
7) Christian Blum, 27. November 2015,
   Self-organization in networks of mobile sensor nodes

Dieser Jahresbericht orientiert sich am Kern des Abschlussberichtes des Graduiertenkollegs METRIK, der im Spätherbst 2015 der Deutschen Forschungsgemeinschaft übergeben worden ist. In die Darstellung der Promotions-/Forschungsprojekte wurden aber nicht mehr die vor 2014 abgeschlossen Arbeiten aufgenommen.

Seit Oktober 2015 befindet sich das Graduiertenkolleg in seiner planmäßigen und finanziell abgesicherten Auslaufphase, die Ende März 2016 ihren Abschluss findet.

Die Aktivitäten im Überblick

Im Graduiertenkolleg METRIK „Modellbasierte Entwicklung von Technologien für selbstoRganisierende Informationssysteme im Katastrophenmanagement, erforschten Informatiker und Geo-Wissenschaftler verschiedener Institute gemeinsam anwendungsorientiert eine besondere Art der drahtlosen Kommunikation und Kooperation von Computern sowie die zielgerichtete Analyse erfasster Daten.

Durch Verwirklichung von Prinzipien der Selbst-Organisation und durch Einsatz preiswerter (aber flächendeckender) Sensorik wurden durch das Graduiertenkolleg neue Horizonte bei der Entwicklung geo-spezifischer Monitoring-, Informations- und Alarmierungssysteme eröffnet, die Umweltprozesse in Raum und Zeit analysieren. Die erforschten Netzarchitekturen zeichneten sich dadurch aus, dass sie ohne eine (aufwändige) zentrale Verwaltung auskommen und sich an Veränderungen der Umgebung anpassen können. Sowohl die Erweiterung solcher Netze als auch ein begrenzter Ausfall von Rechnern und Sensoren schränkte die Arbeitsfähigkeit des Gesamtsystems nicht grundsätzlich ein. Dabei hatten die dafür zu entwickelnden IT-Technologien insbesondere im Kontext eines Katastrophenmanagements nicht nur Anforderungen hinsichtlich ihrer funktionalen Korrektheit, sondern auch hinsichtlich ihrer Zuverlässigkeit und Reaktionsgeschwindigkeit zu berücksichtigen.

Nach einer ersten Phase des Graduiertenkollegs (2006 bis 2010) konnte eine Reihe der bis dahin entwickelten METRIK-Technologien über weitere METRIK-begleitende Projekte mit dem Deutschen GeoForschungszentrum Potsdam, gefördert durch die EU und das BMBF, in prototypische Monitoring-Systeme umgesetzt werden. Eine der überzeugenden Anwendungen der interdisziplinären Zusammenarbeit stellte die modellbasierte Entwicklung des Prototyps eines neuartigen Erdbebenfrühwarnsystems für die stark von seismischen Aktivitäten bedrohte Region Istanbul dar, welcher in der zweiten Phase vom Geoforschungszentrum in weiteren asiatischen Ländern in Minimalkonfigurationen erfolgreich zum Gebäude- bzw. Brückenmonitoring eingesetzt worden ist.

In der zweiten Phase (ab 2010) wurden die METRIK-Technologien konsolidiert und erweitert. Aspekte der Sicherheit, der Mobilität von Sensoren in Gestalt drahtlos kommunizierender Flugroboter, der dynamischen Anpassung des Systems hinsichtlich des momentanen Datenverkehrs durch intelligente Wechsel von Frequenzbändern für die Datenübertragung und eingesetzter Kommunikations- und Abstimmungsregeln beim kooperativen Zusammenwirken der Diensterbringer kamen hinzu. Schließlich wurden Problemlösungen für den Umgang mit großen Mengen erfasster Sensordaten notwendig. Flächendeckend sollten so Daten mittels spezifischer Sensorik zur Erfassung von Temperatur, Luftfeuchte, Verschmutzung, Verkehrsdichte, Energiebedarf, radioaktiver Belastung usw. erfasst und mit anderen geospezifischen Daten abgeglichen werden, die für die Überwachung und Beeinflussung von Umwelt, Gesundheit, Verkehr, Sicherheit und Entwicklung einer städtischen Metropole nicht nur in Extremsituationen von Bedeutung sind. Aus diesem Grund verzahnte sich das Graduiertenkolleg zusätzlich mit der DFG-Forschergruppe Stratosphere, um zu untersuchen, wie das komplexe Informationsdatenmanagement mithilfe dynamisch gebildeter Computer-Cluster (Clouds) noch besser umgesetzt werden kann.

Um nun Umweltprozesse insbesondere auch für die Metropole Berlin prototypisch untersuchen zu können, wurde vom Graduiertenkolleg mit Beginn des Jahres 2011 ein Testnetz am Campus Adlershof der Humboldt-Universität zu Berlin als drahtloses Maschennetzwerk mit 120 Indoor- und Outdoor-Knotenrechner aufgebaut. So bildete in der zweiten Phase die Untersuchung von Grundproblemen einer dynamischen Erfassung und Aggregation von Raum-Zeit-Messungen sowie einer automatischen Ereignisanalyse von Datenströmen einen neuen Schwerpunkt.

Spezifik der zweiten Förderphase

Die gegebenen Hinweise und Empfehlungen aus der Begutachtung der ersten Phase wurden folgendermaßen umgesetzt:

Die empfohlene noch stärkere Konzentration auf Anwendungsprobleme bzw. Bearbeitung domänenspezifischer Aspekte schlug sich hauptsächlich in der Identifikation und Vergabe der einzelnen Aufgabenstellungen sowie der Betreuerzuordnung nieder. Von den 51 bearbeiteten bzw. noch in Bearbeitung befindlichen Themen haben 30 einen stark fokussierten Anwendungsbezug. Die übrigen Arbeiten beziehen zumindest ihre Motivation aus dem praktischen Umfeld von METRIK. Als entscheidende Anwendungsdomänen der interdisziplinären Ausrichtung von METRIK lassen sich die Gebiete Seismologie, Katastrophenschutz, Landnutzung, Geo-Daten-Analyse, Geo-Daten-Integration, Kartographie, Metawerkzeug-Entwicklung, Robotik, Verkehrssteuerung, drahtlose Nachrichtenübertragung und Nano-Optik identifizieren.

Nur als teilweise gelungen sehen wir die geforderte stärkere Berücksichtigung von IT-Sicherheitsaspekten bei der Vergabe von Forschungsthemen an. Im Bereich der Maschennetzwerke im Kontext der Arbeiten am Erdbebenfrühwarnsystem wurden zwei einschlägige sicherheitsorientierte Promotionsprojekte initiiert, deren Bearbeiter aber nach vielversprechender Einarbeitung von der Industrie erfolgreich abgeworben worden sind. So verblieb nur eine aktuelle Arbeit in METRIK. Mit ihr wird der Schutz der Privatsphäre bei Datenbank-Anfrage-Prozessen untersucht. Darüber hinaus wurde aber auch grundsätzlich die deutlichere Betonung von Security-Anforderungen in allen Arbeiten eingefordert. Nachhaltig umgesetzt werden konnte dies „nur“ in Bezug auf Robustheits- und Ausfallsicherheitsaspekte der entwickelten Technologien.

Der Qualitätsbewertung von gesammelten Daten im Geo-Kontext wurden vier Arbeiten gewidmet, von denen zwei von den Bearbeitern vorzeitig abgebrochen worden sind.

Zur gewünschten Förderung der Chancengleichheit sind im Graduiertenkolleg eine Reihe von Maßnahmen ergriffen worden. Zunächst wurden 2010/2011 zwei Junior-Professorinnen (Informatik und Geographie) integriert. Angesichts der hohen

Anzahl von Kindern der METRIK-Kollegiatinnen und Kollegiaten (inzwischen mehr als 50) erwies sich die Anschaffung von interessantem nicht alltäglichem Spielzeug (Lego-Robotik) und Standard-Spielzeug für Kleinstkinder als sehr wirkungsvoll und wurde zur gemeinschaftlich oder individuell organisierten Kinderbetreuung vor Ort bei spontan auftretenden oder regelmäßigen Problemen der METRIK-Eltern überaus gern in Anspruch genommen. Vereinzelt wurden auch Möglichkeiten einer Finanzierungsunterstützung für externe Kinderbetreuung wahrgenommen. Auch die angebotenen Spezial-Trainingsprogramme (Coaching, Mentoring) für die METRIK-Doktorandinnen wurden als sehr hilfreich angesehen und im vollen Umfange genutzt. Darüber hinaus hat sich METRIK aktiv in verschiedene Förderungsinitiativen der Math.-Naturwissenschaftlichen Fakultät der Humboldt Universität zur Erhöhung der Präsenz von Frauen in den Naturwissenschaften am Campus Adlershof eingebracht. Highlights bildeten turnusmäßig stattfindende Workshops zu diesem Thema mit starker METRIK-Beteiligung. Drei der vier METRIK-Kollegiatinnen haben ihre Arbeiten fertig stellen können und das mit sehr guten Ergebnissen.

Forschungsprogramm (2006 – 2016)

Initiale thematische Ausrichtung (erste Förderphase): Da Naturkatastrophen die Menschheitsentwicklung kontinuierlich begleitet haben, sind immer wieder Anstrengungen unternommen worden, den durch diese Katastrophen verursachten Zerstörungen und dem Verlust an Menschenleben entgegenzuwirken. Dabei hat die zuverlässige Vorhersage bzw. schnelle Erfassung von extremen Umweltsituationen (Erdbeben, Stürme und Überschwemmungen) und die zielgerichtete Reaktion darauf immer mehr an Bedeutung gewonnen. Mit den heutigen Entwicklungen in Hardware und Software stehen neue Möglichkeiten zur Verfügung, Katastrophen wirkungsvoll zu begegnen. Der Erschließung dieser Möglichkeiten widmete sich das interdisziplinär ausgerichtete Graduiertenkolleg METRIK „Modellbasierte Entwicklung von Technologien für selbstorganisierende, dezentrale Informationssysteme im Katastrophenmanagement“. In enger Kooperation entwickelten Informatiker und Geo-Wissenschaftler IT-basierte Methoden und Technologien, um den komplexen Prozess des Katastrophenmanagements zumindest in ausgewählten Aspekten zu verbessern.

Im ersten Förderzeitraum konzentrierte sich das Graduiertenkolleg auf die Erforschung, Bereitstellung und Nutzung von Diensten, erbracht durch dynamische, hochflexible und selbstorganisierende Informationssysteme. Der Schwerpunkt lag dabei

a) auf dem Einsatz drahtlos vermaschter Sensornetzwerke
   und deren Integration in bestehende Infrastrukturen,

b) in der Identifikation und Bearbeitung von Problemen bei der Erfassung und Verarbeitung massenhafter (Spatial-)Sensordaten sowie
c) in der Entwicklung von Methoden und Werkzeugen für Anwendungen dezentraler Informationssysteme im Geo-Wissenschaftsbereich unter Einsatz geeigneter Metamodellierungstechnologien.

Das interdisziplinäre Forschungsprogramm wurde durch insgesamt 5 miteinander verwobene Forschungscluster repräsentiert (s. Abb.1).

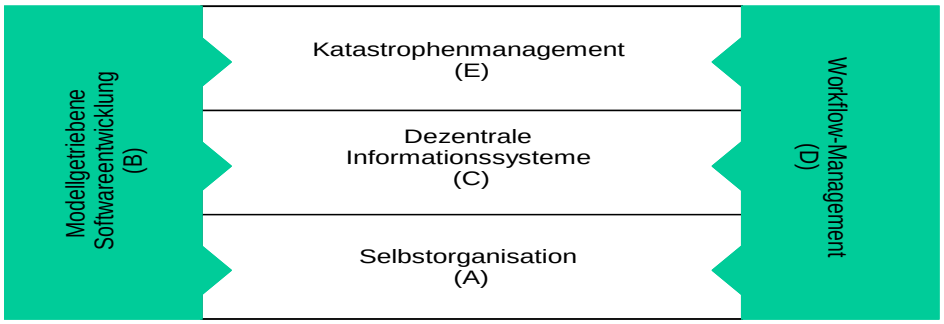


Abb. 1: Cluster des METRIK-Forschungsprogramms

Zu Beginn des Graduiertenkollegs konzentrierten sich die Forschungen insbesondere auf die Funktionen der einzelnen Netzknoten, die für eine Selbstorganisation notwendig sind. Fragestellungen betrafen die kooperative Kommunikation unter schwierigen Bedingungen im Netz (wachsende Netzgröße, Knotenausfall, Signal-Interferenzen), die Replikation dezentraler Datenbestände, ein automatisiertes Deployment und Update von Softwarekomponenten bei laufendem Netzbetrieb, sowie die dynamische Lastverteilung bei Einsatz technisch und energetisch beschränkter Endgeräte. Zudem wurden auch nicht-funktionale Aspekte wie Latenz, und Ausfallsicherheit berücksichtigt.

Aspekte in den weiteren Clustern kamen hinzu. So wurde die Automatisierung des Prozesses *des Findens, der Extraktion und Aggregation relevanter im Internet verbreiteter Informationen unter Berücksichtigung ihrer Chronologie* mit Entwicklung prototypischer Systeme vorangetrieben, wobei Verfahren des Musterabgleichs, der Verarbeitung natürlicher Sprache und des maschinellen Lernens zum Einsatz kamen.

METRIK konnte damit zunächst zum tieferen Verständnis drahtlos vermaschter Computerstrukturen beitragen, angefangen bei ihrer grundsätzlichen Wirkungsweise und Entwicklung bis hin zu ihrer Anwendung und Einsatzoptimierung insbesondere im Kontext eines Katastrophenmanagements.

Abgerundet wurden die Forschungen der ersten Förderphase durch die Entwicklung von Technologien im Metamodellierungsbereich zur Entwicklung domänspezifischer Modellierungssprachen und geeigneter Werkzeuge, die im Geo-Kontext erprobt werden konnten. Untersuchungen grundsätzlicher Art zur

Modellierung von Workflows und ihrer speziellen Ausprägung als Scientific Workflows und die Erweiterung der Anwendungsdomänen leiteten den Übergang in die nächste Phase ein.

Vertiefung, Konsolidierung und Erweiterung des Forschungsprogramms (zweite Förderphase): Die zunächst für die initiale Phase definierten Cluster Selbstorganisation (A), Modellgetriebene Softwareentwicklung (B), Dezentrale Informationssysteme (C), Workflow-Management (D) und Katastrophenmanagement (E) erwiesen sich als außerordentlich tragfähig und wurden deshalb auch für die zweite Förderphase übernommen. Jedoch wurde eine noch engere Verzahnung untereinander angestrebt. Die betreffenden Fragestellungen wurden vertieft und erweitert. Während in der initialen Phase sich die meisten Promotionsprojekte genau einem Cluster zuordnen ließen, waren dies in der Konsolidierungsphase immer mehrere Cluster gleichzeitig.

Zunächst wurden die bislang entwickelten METRIK-Technologien konsolidiert und erweitert. Aspekte der Sicherheit, der Mobilität von Sensoren in Gestalt drahtlos kommunizierender Flugroboter, der dynamischen Anpassung des Systems hinsichtlich des momentanen Datenverkehrs durch intelligente Wechsel von Frequenzbändern für die Datenübertragung und eingesetzter Kommunikations- und Abstimmungsregeln beim kooperativen Zusammenwirken der Diensterbringer kamen hinzu. Schließlich wurden Problemlösungen für den Umgang mit großen Mengen erfasster Sensordaten notwendig. Flächendeckend wurden so Daten mittels spezifischer Sensorik zur Erfassung von Temperatur, Luftfeuchte, Verschmutzung, Verkehrsdichte u.ä. Belastungen mit dem Ziel erfasst, sie mit anderen geospezifischen Daten abgleichen zu können, die für die Überwachung und Beeinflussung von Umwelt, Gesundheit, Verkehr, Sicherheit und Entwicklung einer städtischen Metropole nicht nur in Extremsituationen von Bedeutung sind. Aus diesem Grund verzahnte sich das Graduiertenkolleg zusätzlich mit der DFG-Forschergruppe Stratosphere, um zu untersuchen, wie das komplexe Informationsdatenmanagement mithilfe dynamisch gebildeter Computer-Cluster (Clouds) noch besser umgesetzt werden kann.

Um nun Umweltprozesse insbesondere auch für die Metropole Berlin prototypisch untersuchen zu können, wurde vom Graduiertenkolleg mit Beginn des Jahres 2011 ein Testnetz am Campus Adlershof der Humboldt-Universität als drahtloses Maschennetzwerk mit 120 Indoor- und Outdoor-Knotenrechner aufgebaut. So bildete die Untersuchung von Grundproblemen einer dynamischen Erfassung und Aggregation von Raum-Zeit-Messungen sowie einer automatischen Ereignisanalyse von Datenströmen einen neuen Schwerpunkt des Graduiertenkollegs METRIK. Dabei wurde nach Methoden gefragt, die das Auffinden von Instanzen von Ereignismustern in Ereignisströmen, das Erlernen der Ereignismuster aus

historischen Ereignissequenzen und eine verteilte und parallele Datenstromverarbeitung ermöglichen.

Maßnahmen zur Umsetzung des FO-Programms: Rückblickend erwies sich folgender Maßnahmenkatalog als zielführend und gewinnbringend:

1. Klares Motivationsbild: Zwei Leitmotive spielten in der gesamten Förderphase eine zentrale Rolle. Zum einen trug der interdisziplinäre Charakter des Graduiertenkollegs durch die Kooperation sowohl der Doktoranden als auch ihrer Betreuer aus der Informatik und aus geo-wissenschaftlichem Kontext außerordentlich zur Motivation und dem letztendlichen Erfolg nahezu aller Beteiligten bei. Zum anderen hatte die angestrebte Fokussierung auf modellbasierte Herangehensweisen im Graduiertenkolleg das fachübergreifende Verständnis in einem sehr hohen Maße gefördert.
2. Wertschätzung an der Fakultät und den beteiligten Instituten: Einer örtlich konzentrierten Unterbringung des Graduiertenkollegs wurde durch das Institut für Informatik ein hoher Stellenwert bei der Raumzuteilung, aber auch bei der rechentechnischen Infrastruktrurversorgung beigemessen.
3. Ausbau der bestehenden Kooperation: Ständige Suche der beteiligten Hochschullehrer nach flankierenden Projekten zum Graduiertenkolleg mit Aufnahme assoziierter Doktorandinnen und Doktoranden in das Graduiertenkolleg.
4. Angebot einer lukrativen realen Testumgebung: Aufbau des Humboldt Wireless Labs als Testumgebung für verschiedene Promotionsarbeiten (ermöglicht durch vorherige Industriekooperationen zur Produktion der benötigten Knotenrechner mit ausgestatteter Sensorik) und Unterstützung durch studentische Hilfskräfte. So konnte ein 120-Knoten-Netz in Berlin aufgebaut werden und die Versorgung des Kooperationspartners Deutsches GeoForschungszentrum Potsdam beim Aufbau von seismischen Monitoring-Infrastrukturen im asiatischen Raum gesichert werden.
5. Verallgemeinerung: Die Suche nach weiteren Anwendungsdomänen bei der Entwicklung domänenspezifischer Modellierungsprachen am Beispiel einer erfolgreichen Kooperation mit dem Forschungsbereich NanoOptik (Institut für Physik).
6. Betreuer-Struktur: Verjüngung und Erweiterung des Teams betreuender Hochschullehrer und des Spektrums beteiligter Partnereinrichtungen zu Beginn der zweiten Förderperiode. Die bestehende erfolgreiche Kooperation mit der TU Eindhoven zum Abschluss von Doppelpromotionen (Cotutelle de thèse) im Bereich der Workflow-Modellierung wurde fortgesetzt.

Technologischer und wissenschaftlicher Gesamtertrag der 10-jährigen Förderperiode:

Im Zeitraum der ersten Förderperiode wurden folgende Akzente gesetzt:

(1) Neben wertvollen wissenschaftlichen Publikationen konnte auch die erfolgreiche Nutzung von METRIK-Technologien bei verschiedenen prototypischen

Anwendungen im Geo-Kontext unter Beweis gestellt werden. Insbesondere ist es gelungen, in Experimenten einen zentralen Anspruch zu untermauern: Dieser besteht darin, dass selbstorganisierende Systeme in Gestalt drahtlos vermaschter Sensornetzwerke durch Miniaturisierung, autonome Adaptionsfähigkeit und niedrige Kosten völlig neue Möglichkeiten eröffnen, komplexe Umweltprozesse zu vermessen und in Echtzeit zu analysieren. So wurde in Kooperation mit dem europäischen Projekt SAFER ein Erdbebenfrühwarn- und Rapid-Response-System für die Stadt Istanbul entwickelt (s. Abb.2), dessen Netzwerkinfrastruktur mit solchen von METRIK entwickelten Algorithmen betrieben wird. Dies ist das weltweit erste Erdbebenfrühwarn- und Rapid-Response-System, das in einer dicht besiedelten Metropole mit Messtechnik vor Ort zum prototypischen Einsatz kommt. Dies wurde nur durch den Einsatz preiswerter drahtlos vermaschter, selbstorganisierender Netzwerke möglich. Weitere Erfahrungen konnten durch die Anwendung der METRIK Technologie zur Gebäudeüberwachung im Zuge der intensiven Serie von Nachbeben des L'Aquila-Ereignisses (April 2009) in Italien gesammelt werden. Inzwischen befinden sich derartige Netzkonfigurationen kleinerer Größe in vielen asiatischen Ländern im Einsatz, mit denen das Deutsche GeoForschungszentrum Potsdam Projektkontakte unterhält.

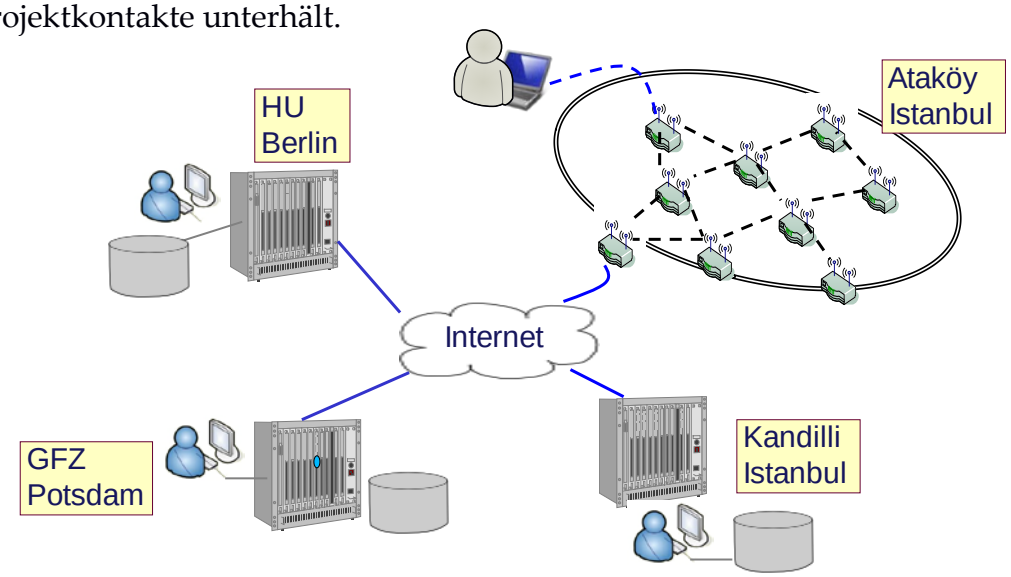


Abb. 2: Erdbebenfrühwarn- und Rapid-Response-System für Istanbul

(2) Die Planung und Durchführung des L'Aquila-Einsatzes der deutschen Erdbeben Task Force (eine Gruppe an diesem Institut GFZ) wurde durch das in METRIK entwickelte Informationssystem EQUATOR unterstützt, das dem Team eine integrierte und ständig aktualisierte Sicht auf weltweit verteilte Erdbebeninformationen ermöglichte.

(3) Weitere, sehr ermutigende Forschungsresultate von METRIK liegen bei der Workflow-basierten Planung von Katastropheneinsätzen vor; diese wurden

ebenfalls in enger Kooperation mit der o.g. Erdbeben Task Force entwickelt und unterlagen dort einer stetigen Evaluation.

Diese Beispiele zeigen, dass sich das METRIK-Kolleg im Bereich der IT-Ausgestaltung geo-spezifischer Forschungsprojekte in Berlin-Brandenburg als wichtiger Expertise-Träger in der ersten Förderphase etablieren konnte. Sie zeigen außerdem, wie dem zentralen Anspruch des Kollegs hinsichtlich einer interdisziplinären Methodenentwicklung entsprochen werden konnte.

Bei der Konsolidierung selbstorganisierender Technologien für drahtlose Sensor-/Aktornetze konzentrierte sich das Graduiertenkolleg in der zweiten Förderperiode auf Probleme des WLAN-Standards IEEE 802.11 und suchte dabei insbesondere spezielle Herausforderungen derartiger Infrastrukturen im Kontext künftiger intelligenter Städte (Smart City). Neuartige Anwendungen wie in Bereichen des Umweltmonitorings oder des Verkehrsmanagements lassen sehr große Mengen an Daten und eine hohe Diversität der Datentypen erwarten, auf deren Basis Prozesse für eine Vielzahl von Teilnehmern in Echtzeit zu koordinieren sind. Daraus ergeben sich hohe Anforderungen in Hinblick auf Durchsatz, Latenz, Zuverlässigkeit sowie Energieeffizienz. Insbesondere kommt der Erfassung und Auswertung interner sowie externer Interferenzen eine wichtige Rolle zu, da die Datenübertragungen im lizenz-freien Spektrum ablaufen.

(1) Konkret wurden hier neuartige Verfahren auf der Mediumzugriffsschicht entwickelt, die die hohen Anforderungen an E2E Zuverlässigkeit, Latenz und Durchsatz auch im Falle von signifikanter externer Interferenz erfüllen. Die entscheidende Idee bestand darin, die Relay- bzw. Pfad-Diversität auszunutzen. Im Gegensatz zu traditionellen Verfahren werden hier keine festen Pfade zwischen den Kommunikationsteilnehmern verwendet. Stattdessen leitet man die Pakete je nach Interferenzsituation und Kanalschwund (fading) adaptiv über verschiedene Relays/Pfade um.

(2) Weitere Untersuchungen hinsichtlich der Zuverlässigkeit der Nachrichtenzustellung zeigten, dass herkömmliche Methoden zur Optimierung von Network Wide Broadcasts in der Realität in Bezug auf die Zuverlässigkeit oft scheitern und nur eine schlechte Verteilung einer Nachricht in einem drahtlosen Maschennetzwerk erreichen. So wurden Verfahren entwickelt, die die Zuverlässigkeit erhöhen, wobei aktuelle Medienzugriffsverfahren bezogen auf das Broadcast Storm Problem besser als erwartet arbeiten. Bei Nachrichten (wie Alarm-Auslöser) mit hoher Zustellungspriorität aber niedriger Frequenz, sollte man daher in typischen Maschennetzwerken möglichst keine Redundanzoptimierung durchführen, sondern besser noch in Kombination mit sinnvollen Bestätigungsverfahren die Redundanz erhöhen.

(3) Ein weiterer untersuchter Ansatz für den Umgang mit Interferenz in drahtlosen Maschennetzen ist die verteilte Kanalzuweisung. Hier werden sich nicht-überlappende Kanäle im verfügbaren Frequenzspektrum für Übertragungen verwendet, die auf dem gleichen Kanal Interferenzen erzeugen würden. Dieser Ansatz ist möglich, da die verwendeten Funktechnologien, wie zum Beispiel IEEE 802.11 (WLAN), mehrere nicht-überlappende Kanäle bereitstellen. Aufgrund der großen Verbreitung dieser Technologie, ist eine hohe Dichte von privaten wie kommerziellen Netzen im urbanen Raum die Norm. Diese räumlich überlappenden Netze konkurrieren um den Medienzugriff und die dadurch entstehende Interferenz kann die Netzleistung mindern. Daher ist es für die Leistung von Kanalzuweisungsalgorithmen von großer Bedeutung, die Aktivität der externen Netze mit einzubeziehen. Im Rahmen von METRIK wurde die Entwicklung eines messungs-basierten Interferenzmodells vorangetrieben, mit dem Interferenzabhängigkeiten der Maschenrouter untereinander effizient bestimmt werden können. Weiterhin wurde ein Algorithmus für die verteilte Kanalzuweisung entwickelt, der die Aktivität von externen Netzen berücksichtigt.

(4) Die speziell entwickelten Protokolle wurden im Humboldt Wireless Lab ausführlich untersucht und konnten ihre Praktikabilität eindrucksvoll unter Beweis stellen, wozu vorab die Bereitstellung einer geeigneten Experimentierplattform notwendig wurde, die mit der Größe der Datenmengen umgehen kann. Zudem musste angenommen werden, dass diese Daten heterogen sind. Je nach Ursprung (z. B. Art, Typ, Version des Sensors oder Kontext, Ort, Zeit der Messung) sind diese Daten von unterschiedlicher Struktur (Syntax) und haben eine unterschiedliche Bedeutung (Semantik). Software zur automatischen Analyse dieser Daten (wie in METRIK bereitgestellt) muss diesem Umstand gerecht werden und Daten unterschiedlicher Art und Herkunft auch unterschiedlich behandeln, auch wenn diese Unterschiede oft nur einzelne Aspekte der Datenverarbeitung betreffen. Mit Techniken der modellbasierten Softwareentwicklung gelang es, durch geeignete Abstraktion den generischen Anteil der Software vom spezifischen zu trennen und dadurch die wiederholte Implementierung von großen Teilen der Datenprozessierung zu vermeiden und eine effizientere und sichere Entwicklung datenverarbeitender Software für Smart City-Anwendungen zu erlauben. In METRIK wurde der Ansatz verfolgt, zwischen der Datenrepräsentation in modellbasierten Entwicklungstechnologien und der Datenrepräsentation in existierenden Technologien zur Verarbeitung großer Datenmengen (Datenbanken, verteilte Datenverarbeitung) Abbildungen zu finden, welche genutzt werden können, um beide technologischen Domänen miteinander zu verknüpfen und so die Fähigkeit sehr große Datenmengen zu verarbeiten auch der modellbasierten Softwareentwicklung zur Verfügung zu stellen. So entstand ein Framework zur

Persistierung von Modellen in Dokumenten- oder Spalten-orientierten NoSQL-Datenbanken, wie sie zum Beispiel zur verteilten Datenspeicherung im Cloud-Computing verwendet werden.

(5) Aufgrund des Umfangs der erfassten und zu analysierenden Daten ist i.allg. eine Verarbeitung auf einem einzelnen Rechner nicht mehr möglich. So kam eine verteilte Datenstromverarbeitung auf den Plan. Entsprechende Systeme stellen die Forschung jedoch vor neue Herausforderungen und Fragestellungen. Nach dem aktuellen Stand der Technik werden verteilte Datenstrom-verarbeitungssysteme manuell aufgesetzt und optimiert. Unter anderem ist dabei die Clustergröße, also die Anzahl der verwendeten Rechner, ein maßgeblicher Kostenfaktor. Werden zu wenig Rechner eingesetzt, kann der Datenstrom nicht zeitnah verarbeitet werden, was entweder zu hohen Latenzen oder sogar zu Datenverlust führen kann. Gleichzeitig führt eine zu großzügige Bereitstellung von Rechnern zu unnötigen Mehrkosten. Des Weiteren ändern sich die Anforderungen an ein Datenstromverarbeitungssystem über die Zeit (oft zyklisch). Für diesen Fall ist es wünschenswert, die Clustergröße dynamisch anzupassen. Deshalb werden Datenstromverarbeitungssysteme auch oft in einer „Cloud“ betrieben, die dynamische Skalierungen ermöglicht. METRIK zeigte, wie Kostenmodelle für die verteilte Datenstromverarbeitung zur Abschätzung benötigter Rechenkapazitäten entwickelt werden können, um den Betrieb von verteilten Datenstromverarbeitungssystemen zu erleichtern. Dabei müssen mehrere Kostenfaktoren wie CPU-Verbrauch, Hauptspeicherverbrauch und Netzwerkkosten abgeschätzt werden. Gleichzeitig müssen die einzelnen Rechenoperationen effizient auf die zur Verfügung stehenden Ressourcen verteilt werden, wozu geeignete Scheduling-Verfahren benötigt werden.

(6) Ein weiteres wichtiges Forschungsergebnis ist mit der automatisierten Analyse von Ereignisströmen in unterschiedlichen Domänen verbunden, die auf das Auffinden von Instanzen von Ereignismustern in Ereignisströmen, das Erlernen der Ereignismuster aus historischen Ereignissequenzen und die verteilte und parallele Datenstromverarbeitung ausgerichtet sind. Ereignismuster beschreiben dabei zusammengesetzte Ereignisse und beinhalten Bedingungen hinsichtlich der zeitlichen Ordnung, der Attributwerte sowie der Anzahl und der zeitlichen Ausdehnung jener Ereignisse, die gemeinsam das zusammengesetzte Ereignis darstellen.

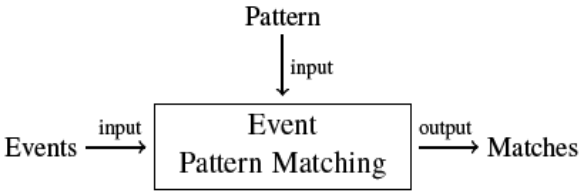


Abb. 3: Event Pattern Matching (EPM)

Hierzu wurden in der zweiten Förderperiode eine Vielzahl von Beiträgen verschiedener Kolleg-Mitglieder geleistet, die überwiegend als kooperative Publikationsleistungen entstanden sind und den Einzelberichten zu entnehmen sind.

(7) Mit der Einbeziehung der Mobilität von Sensoren in Gestalt drahtlos kommunizierender Flugroboter wurde in METRIK u.a. die Frage nach einer geeigneten Kontrollstrategie für autonome Roboter aufgeworfen, die die Roboter in die Lage versetzt, sich in drahtlosen (ad hoc) Multihop-Netzwerken integrieren zu können. Die vorgeschlagene Kontrollstrategie setzt sich aus lokalen Strategien der einzelnen Roboter zusammen und koppelt so die einzelnen Roboter lose durch sensomotorische Schleifen. Diese lose Kopplung führt schließlich zu einer globalen Selbstorganisation des Systems. In METRIK wurde ein Algorithmus zur Exploration unbekannter Netzwerke für großflächige Außeneinsätze entwickelt und per Simulation evaluiert, wobei sich der Algorithmus als sehr tolerant gegenüber Messrauschen erwiesen hat. Für den Fall mehrerer Roboter, d.h., als Schwarm drahtlos verbundener Roboter, wurden bereits viele interessante Algorithmen in der Literatur vorgestellt. Allerdings wird im Moment keiner dieser Algorithmen in realen Szenarien eingesetzt. Einer der Gründe hierfür ist, dass naive Schwarmalgorithmen meist nicht sicher genug und fehlertolerant sind. Um dieses Problem zu lösen, wurde der Einsatz einer Architektur basierend auf internen Modellen, die einen internen Simulator nutzt, untersucht, um die Sicherheit und die Fehlertoleranz dieser Systeme zu erhöhen und sie damit geeigneter für reale Anwendung zu machen. Die vorgeschlagene Architektur hat sich in diesen Experimenten als sehr effektiv erwiesen.

(8) Ein weiteres Problemfeld betrifft die effiziente Analyse von sehr langen Zeitreihen, die im Realfall tatsächlich im Deutschen GeoForschungszentrum Potsdam anfallen. Die Untersuchungen hoch-dimensionaler Daten erfolgen hier am Beispiel der sogenannten Recurrence Quantification Analysis (RQA), einer Methode aus der Zeitreihenanalyse. Konkret werden hier das Problem der effizienten Berechnung langer Zeitreihen sowie der effizienten Exploration großer Ergebnismengen adressiert. Um das Problem der Verarbeitung sehr langer Zeitreihen behandeln zu können, wurde zunächst ein Ansatz zur Partitionierung der Recurrence Matrix entwickelt, der auf dem Konzept Divide

& Recombine basiert. Ansätze zur effizienten Durchführung der RQA werden in zwei Richtungen untersucht: Die Parallelisierung des RQA-Verfahrens und der Einsatz von Baumstrukturen, um die Anzahl der durchzuführenden Ähnlichkeitsvergleiche drastisch zu reduzieren.

Promotionsprojekte und bisher erzielte Forschungsergebnisse

BJÖRN LICHTBLAU

Reliable Network-Wide Broadcasts for Wireless Mesh Networks

Drahtlose Maschennetzwerke (Wireless Mesh Networks, WMNs) sind selbstorganisierende Netzwerke die aus vielen drahtlos kommunizierenden Geräten bestehen. Sie können ad hoc ausgebracht werden und kommen ohne weitere Infrastruktur aus. Damit sind sie für viele Anwendungen interessant, u.a. sind sie für die spontane oder dauerhafte Nutzung als Frühwarnsystem im Katastrophenmanagement oder im Kontext von intelligenten Städten geeignet. Eine grundlegende Operation solcher Netzwerke ist ein netzwerkweiter Broadcast (NWB), mit der eine Nachricht von einem Knoten an alle anderen Knoten im Netzwerk übertragen wird. Netzwerkweite Broadcasts werden üblicherweise von Wegewahlverfahren benötigt und eingesetzt. Ein anderer Anwendungsfall ist die Verteilung von Warnnachrichten nach der Detektion einer kritischen Situation zur Alarmierung oder automatisierten Reaktion wie es z.B. in einem Frühwarnsystem nötig wäre. Hierbei kommt es vorrangig auf die Zuverlässigkeit der NWBs an, um alle Netzwerkknoten zuverlässig zu erreichen.

Die triviale Realisierung eines NWBs in WMNs ist das auch aus drahtgebundenen Netzwerken bekannte Fluten, wobei jeder Knoten beim ersten Empfang eines NWBs diesen genau einmal wiederholt. Da in einem gut vermaschten Netzwerk dadurch viele redundante Nachrichten erzeugt werden die in einem geteilten Medium konkurrierend versendet werden, kann dies in WMNs zu Paketverlusten aufgrund von Kollisionen führen (das sogenannte Broadcast Storm Problem). Daher ist das Ziel der meisten Ansätze, diese Redundanz zu reduzieren oder zu entfernen. Dabei wird aber fast immer die Zuverlässigkeit außer Acht gelassen.

Diese Arbeit untersucht und entwickelt deshalb Verfahren um NWBs mit hoher Zuverlässigkeit zu realisieren, um sie für seltene aber wichtige Nachrichten einsetzbar zu machen. Hierfür werden lokale Informationen, wie die Übertragungsqualität zu den Nachbarn oder die Dichte der Nachbarschaft im Maschennetzwerk genutzt, um redundante Nachrichten zu reduzieren und gleichzeitig eine hohen Nutzen für jede Übertragung zu erreichen. Dabei wird im Gegensatz zu existierenden Ansätzen ein probabilistisches gewichtetes Graphmodell zugrunde gelegt, wobei die Übertragungsqualitäten den Kantengewichten entsprechen. Basierend auf diesem probabilistischem Graphenmodell wurde weiterhin ein auf

einer Monte-Carlo-Simulation basierendes Verfahren entwickelt, mit der sich die Erreichbarkeit von Knoten durch einen NWB effizient approximieren lässt.

Weiter werden Bestätigungen verschiedener Art und gegebenenfalls Übertragungswiederholungen im Kontext von NWBs genutzt um die Verteilung abzusichern. Die Evaluation existierender und neu entwickelter Ansätze erfolgt experimentell und wurde sowohl in Simulationen als auch in real existierenden Testnetzwerken (dem Humboldt Wireless Lab der Humboldt-Universität zu Berlin und dem DES-Testbed der Freien Universität Berlin) durchgeführt. Um die selben Implementierung von NWB-Protokollen sowohl in Simulationen als auch auf echten Maschennetzwerkknoten zu nutzen wurde ein spezielles Framework genutzt und weiterentwickelt.

Zusammengefasst zeigen die interessantesten Ergebnisse, dass herkömmliche Methoden zur Optimierung von NWBs in der Realität im Bezug auf die Zuverlässigkeit oft scheitern und nur eine schlechte Verteilung einer Nachricht in einem drahtlosen Maschennetzwerk erreichen. Die vorgestellten Verfahren können die Zuverlässigkeit erhöhen, wobei aktuelle Medienzugriffsverfahren bezogen auf das Broadcast Storm Problem besser als erwartet arbeiten. Hat die Zuverlässigkeit der Nachrichten absolute Priorität, und ist die Frequenz solcher Nachrichten eher gering (wie z.B. bei einem Alarm), sollte man daher in typischen WMNs sogar gar keine Redundanzoptimierung durchführen, sondern besser noch in Kombination mit sinnvollen Bestätigungsverfahren die Redundanz erhöhen.

Die Dissertation ist zu grossen Teilen fertiggestellt, die Fertigstellung der Dokumentation von Experimentausführung und Ergebnisdiskussion fehlen aber noch und die finale Überarbeitung ist noch offen. Das Dissertationsvorhaben war im Modellteil bezüglich der mathematischen Modellierung mit der Arbeit von Andreas Dittrich aus dem Graduiertenkolleg vernetzt und führte zu gemeinsamen Publikationen, das Verfahren zur effizienten Berechnung der Erreichbarkeit durch NWBs diente dabei als Eingabe für seine Modellierung von Service Discovery Protokollen, die auch mit NWBs arbeiten.

Forschungsaufenthalte, Konferenzen, Workshops:
19th IEEE Symposium on Communications and Vehicular Technology in the Benelux, November 2012.

14th International IEEE Symposium on a World of Wireless, Mobile and Multimedia Networks (WoWMoM), Madrid, Spain, Juni 2013.

Eigene Publikationen:
JOACHIM FISCHER, JENS-PETER REDLICH, JOCHEN ZSCHAU, CLAUS MILKEREIT, MATTEO PICOZZI, BJÖRN LICHTBLAU, INGMAR EVESLAGE, *Erdbebenfrühwarnsystem für Istanbul: Drahtlos kommunizierende Ad-hoc Netzte mit Echtzeit-Sensorinformation.* Bulletin SEV/VSE 4, p.35-39, 2011.

JOACHIM FISCHER, JENS-PETER REDLICH, JOCHEN ZSCHAU, CLAUS MILKEREIT, MATTEO PICOZZI, KEVIN FLEMING, MIHAL BRUMBULLI, BJÖRN LICHTBLAU, INGMAR EVESLAGE, *A wireless mesh sensing network for early warning.* Journal of Network and Computer Applications, 35(2):538-547, March 2012.

BJÖRN LICHTBLAU, JENS-PETER REDLICH, *Network-Wide Broadcasts for Wireless Mesh Networks with Regard to Reliability.* Proceedings of 19th IEEE Symposium on Communications and Vehicular Technology in the Benelux, November 2012.

BJÖRN LICHTBLAU, JENS-PETER REDLICH, *Link Quality Based Forwarder Selection Strategies For Reliable Network-Wide Broadcasts.* (in press) 5th IEEE International Workshop on Hot Topics in Mesh Networking, June 2013.

ANDREAS DITTRICH, BJÖRN LICHTBLAU, RAFAEL RIBEIRO REZENDE, MIROSLAV MALEK, Modeling *Responsiveness of Decentralized Service Discovery in Wireless Mesh Networks*, 17th International GI/ITG Conference on "Measurement, Modelling and Evaluation of Computing Systems" and "Dependability and Fault-Tolerance", Bamberg, Germany, March 17-19, 2014.

BJÖRN LICHTBLAU, ANDREAS DITTRICH, *Probabilistic Breadth-First Search – A Method for Evaluation of Network-Wide Broadcast Protocols*, 6th IEEE/ACM/IFIP International Conference on New Technologies, Mobility and Security, NTMS 2014, Dubai, UAE, March 30 – April 2, 2014.

LARS DÖHLING

Time-aware Information Extraction

Nach einer Katastrophe wie z. B. einem Erdbeben gibt es seitens von Entscheidungsträgern einen großen Informationsbedarf über die Auswirkungen des Ereignisses. Solche Informationen lassen sich im zunehmenden Maße in textueller Form im Internet finden. Dies umfasst sowohl konventionellen Quellen wie Zeitungen als auch moderne wie Blogs und Newsgroups oder sozialen Netze wie Twitter oder Flickr. Diese Quellen bieten mit die detailliertesten verfügbaren Informationen, aber ihre manuelle Auswertung ist eine zeit- und daher kostenaufwändige Aufgabe. Dies gilt besonders unter dem Aspekt einer sich ständig ändernden Informationslage, welche eine kontinuierliche – manuelle – Aktualisierung bedingt. Je nach Art des Ereignisses, der benötigen Information und der verstrichener Zeit können 1 Tage alte Dokumente noch korrekte oder bereits überholte Fakten enthalten.

Im Rahmen meiner Dissertation untersuche ich daher Methoden, welche das Finden, die Extraktion und Aggregation relevanter Informationen unter Berücksichtigung der Zeitdimension automatisieren. Die Methoden beruhen auf Verfahren des Musterabgleichs, der Verarbeitung natürlicher Sprache und des maschinellen Lernens. Dabei konzentriere ich mich auf drei Fragestellungen:

1. Wie exakt lassen sich ereignisrelevante Dokumente zeitlich verorten?

2. Wie und mit welcher Genauigkeit lassen sich relevante Informationen in Form von mehrstelligen Relationen aus natürlich sprachlichen Texten extrahieren?
3. Wie und mit welcher Güte können Extraktionsergebnisse aus verschiedenen Quellen, welche über dasselbe Ereignis zu verschiedenen Zeitpunkten berichten, zu einen kohärenten Gesamtbild fusioniert werden?

Die dafür entwickelten Methoden werden als Module eines konfigurierbaren Frameworks (Abbildung 1) implementiert und anhand von Texten aus dem Bereich Erdbeben und Überflutungen evaluiert.

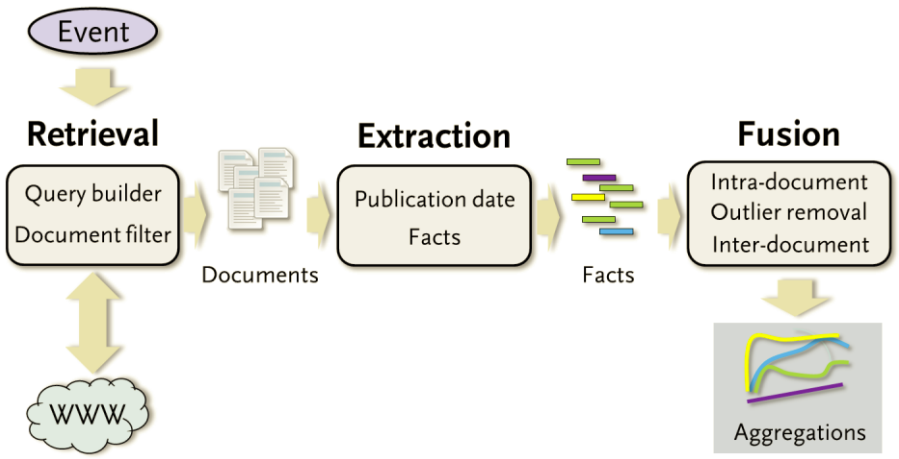


Abbildung 1

Das Retrieval-Modul liefert, basierend auf dem Ereignis, die zu analysierenden Dokumente. Ein Ereignis wird hierbei als ein Triplet aus Typ (z.B. Erdbeben), Zeit (UTC-Zeitmarke) und Ort (GPS-Koordinaten) modelliert. Ausgehend vom Ereignis werden Suchanfragen generiert und an gängige Suchmaschinen übermittelt. Eine stichprobenartige Evaluation für Bing und Google liefert hier Quoten von rund 75% relevanter Suchtreffer (Precision). Das Generieren geeigneter Suchanfragen stellt an sich keinen besonderen wissenschaftlichen Beitrag dar, ist aber für mein Promotionsprojekt zwingend notwendig. Es dient der Dokumentenakquise, welche die Evaluation der übrigen Komponenten erst ermöglicht.

Das Extraction-Modul ordnet die Dokumente (und ihre enthaltenen Fakten) sekundengenau nach ihrem Publikationsdatum. Dabei verfolge ich einen musterbasierten Ansatz, welcher im HTML-Code enthaltenen Zeitstempel verwendet. Besondere Herausforderungen stellen dabei Ambiguitäten dar, welche auf verschiedenen Ebenen auftreten. Dies sind z.B. die unterschiedliche Reihenfolge von Tag, Monat im amerikanischen (MM/dd) und britischen (dd/MM) Englisch oder ambigue Zeitzonenbezeichner wie PST (Pacific Standard Time (UTC−08)

versus Philippine Standard Time (UTC+08)). Eine umfassende Evaluation auf 116 unterschiedlichen HTML-Seiten von verschiedenen Hosts ergibt eine Genauigkeit meines Ansatzes von 31% auf die Sekunde und 47% auf die Minute. Bei einer Toleranz von 12 Stunden steigt dieser Wert auf über 91%. Ein Vergleich mit CarbonDate und DCTFinder, zwei alternative Ansätze zur zeitlichen Verortung, zeigt eine Überlegenheit meines Ansatzes auf den Evaluationsdaten.

Die zweite Aufgabe des Moduls ist die Extraktion der relevanten Fakten aus den Dokumenten. Hierbei betrachte ich Fakten als Tupel n-stellige Relationen. Eine Relation ist eine Menge von geordneten Tupeln, welche eine semantische Beziehung beschreibt. Je nach Tupelgröße spricht man von einstelligen, zweistelligen oder n-stelligen Relationen. Ausgehend von einer Erkennungswahrscheinlichkeit jedes Elementes eines Tupels <1 und jedes semantischen Zusammenhangs zwischen Elementen von ebenfalls <1, ergibt sich vereinfacht, dass die automatische Extraktion u.U. schwieriger wird, je größer die Tupel sind. Die Extraktion erfolgt dabei in zwei Schritten: Zuerst werden die möglichen Elemente eines Tupels identifiziert (Named Entity Recognition, NER) und anschließend Relationen zwischen diesen gebildet (Relationship Extraction, RE). Für die NER verfolge ich zwei Ansätze: Wörterbuch/regulärer Ausdruck und maschinelles Lernen, konkret ein Conditional Random Field (CRF). Für die RE setzte ich sowohl einfache Kookkurrenz als auch Musterabgleich in Dependenzgraphen und maschinelles Lernen ein, hier Supportvektormaschinen (SVMs). Die Evaluation der verschiedenen (Kombinationen von) Methoden erfolgt auf drei neuen, englischsprachigen Korpora in den Domänen Erdbeben und Überflutung. Hierbei werden Opferangaben, modelliert als 4-stellige Relation, extrahiert. Diese Opferangaben, d.h. Anzahl der Verwundeten, Vermissten, Toten, Verschütteten usw., sind eine wichtige Kenngröße zur Abschätzung der Auswirkung einer Katastrophe und damit der benötigten Hilfsressourcen. Die Evaluation zeigt, dass der Ansatz des maschinellen Lernens – CRF plus SVM – dem Wörterbuchansatz in Kombination mit dem Musterabgleich nicht überlegen ist. Beide erreichen für die 4-stellige Relation Werte im Bereich 75% F1, dem harmonische Mittel aus Precision (Korrektheit) und Recall (Vollständigkeit). Experimente mit perfekter NER zeigen, dass die NER der limitierende Faktor in der Extraktionskette ist, nicht die gewählte RE-Methode. In dieser Konfiguration werden F1-Werte von über 90% erzielt. Damit sollte die NER im Fokus zukünftiger Untersuchungen liegen. Die gelernten Extraktionsmodelle, besonders die Dependenzmuster, zeigen sich auch robust gegenüber einem Domänenwechsel, also z.B. von Erdbeben zu Überflutungen. D.h. gelernte Modelle lassen sich auch auf neuen Domänen anwenden, für welche noch keine annotierten Trainingsdaten existieren.

Das Fusion-Modul hat die Aufgabe, die extrahierten Fakten, welche dasselbe Ereignis zu verschiedenen Zeitpunkten und aus der Sicht verschiedener Quellen

beschreibt, zu eine kohärenten Gesamtbild zu vereinen. D.h. für jeden Zeitpunkt nach dem Ereignis wird ein Fakt zurückgegeben. Diese Aggregation wird zusätzlich durch die Fehlerpropagierung innerhalb des Frameworks erschwert: irrelevante Dokumente, eine falsche zeitliche Einordung oder falsch-extrahierte Fakten. Hier schlage ich ein zweistufiges Verfahren aus Intra- und Interdokumentfusion vor. Die Intradokumentfusion liefert für jedes Dokument maximal einen Fakt zurück. Ausgehend von der Struktur eines typischen Newsartikels – die zentralen Fakten werden in der Überschrift und/oder im ersten Absatz genannt – wird dazu die erste gefundene Angabe oder die häufigste Angabe im Artikel gewählt. Auf das Ergebnis der Intradokumentfusion wird nun ein zeitsensitiver Ausreisfilter angewandt, welcher unwahrscheinliche Datenpunkte/Fakten eliminiert. Dessen Ausgabe kann auf verschiedenen Arten in der Zeit aggregiert werden: über das Fitten fakten-spezifischer Kurven, welche „typische“ Werteverläufe modellieren, oder gleitende Mittelwerte. Die Evaluation des Fusions-Moduls erfolgt im Zuge einer Gesamtevaluation des Frameworks. Dazu werden die zeitlichen Verläufe der Opferzahlen von über 30 Erbeben und Flutkatastrophen zwischen 2006 und 2012 betrachtet. Als Gold Standard dienen hierbei die Revisionen der Wikipedia-Info-boxen. Pro Ereignis werden bis zu 100 Suchanfragen an die Bing API gesandt, was im Schnitt zu 1862 Dokumente pro Ereignis führt. Diese werden zeitlich verortet und die Opferangaben mit trainierten Modellen extrahiert. Als Fusionskurve wird eine Hyperbel, eine Sättigungsfunktion, verwendet und mit einem gleitenden arith-metischen Mittel verglichen. Beide Fusionsstrategien ergeben für beide Domänen eine relative Abweichung vom Referenzverlauf von im Schnitt 20%. Abbildung 2 zeigt exemplarisch die Ausgabe der Frameworks („Fitted Curve“ bzw. „Sliding average“) in Relation zum Gold Standard („Wikipedia reference“) für das Yushu Erdbeben in 2010. „Inlier“ beschreibt die Datenpunkte nach dem Ausreißfilter.

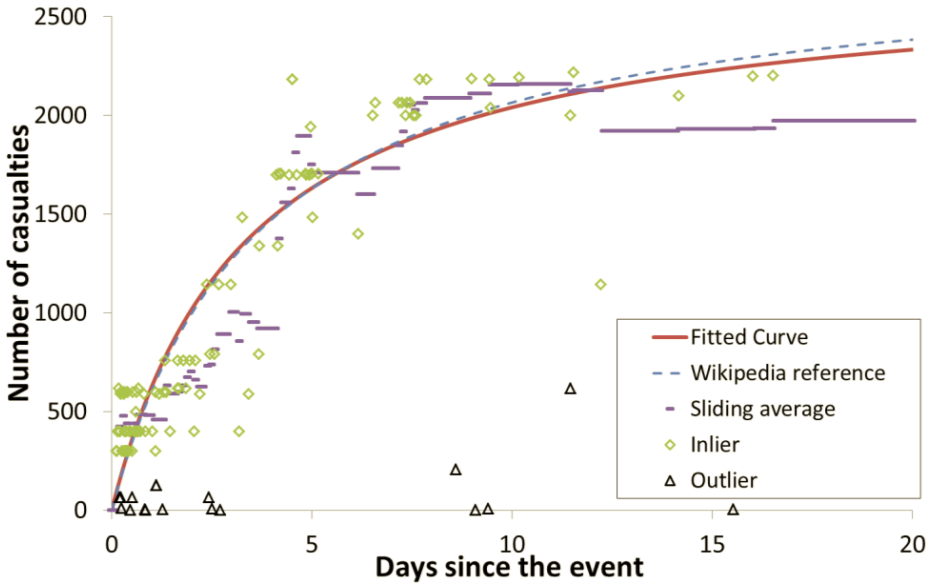


Abbildung 2

Je nach Anwendungsszenario können diese 20% Abweichung einen brauchbaren Wert darstellen. Zum Beispiel möchte man zur Kalkulation benötigter Hilfsressourcen wissen, ob mit 20, 200 oder 2000 Verletzten zu rechnen ist. Der vorgestellte automatische Ansatz könnte hier arbeitserleichternd eingesetzt werden. Betrachtet man zusätzlich die Komplexität der verwendeten Informationsverarbeitungskette, so sind die gemessenen 20% als Erfolg zu werten.

Die Gliederung der anzufertigenden Monographie orientiert sich an den Modulen des Frameworks. Zu jedem Modul existieren bereits schriftliche Ausarbeitungen, teilweise sind diese veröffentlicht. Die Experimente sind abgeschlossen und der aktuelle Fokus meiner Arbeit liegt im Zusammenführen der Teile.

Forschungsaufenthalte, Konferenzen, Workshops:

8th European Summer School on Information Retrieval (ESSIR 2011), Koblenz, August 2011.

10th International Semantic Web Conference (ISWC 2011), Terra Cognita Workshop Bonn, Oktober 2011.

INFORMATIK 2011, Jahrestagung der Gesellschaft für Informatik, Berlin, Oktober 2011.

Informare! 2012, Berlin, Mai 2012.

23rd World Wide Web Conference (WWW 2014), Temporal Web Analytics Workshop, Seoul, Republik Korea, April 2014.

9th Edition of its Language Resources and Evaluation Conference (LREC 2014), Reykjavik, Island, Mai 2014.

fit4sec Workshop 2015.

Eigene Publikationen:

DÖHLING, L., LEWANDOWSKI, J., AND LESER, U., *A Study in Domain-Independent Information Extraction for Disaster Management*, Workshop on Disaster Management and Principled Large- scale information Extraction for and post emergency Logistics (DIMPLE), Reijkjavik, Island, (2014).

DÖHLING, L. AND LESER, U., *Extracting and Aggregating Temporal Events from Text,* 4th Temporal Web Analytics Workshop (TempWeb), Seoul, Korea, (2014).

DÖHLING L., WOITH H., FAHLAND D., AND LESER U., *Equator: Faster Decision Making for Geo-scientists,* GI-Jahrestagung 2011, Workshop on IT support for rescue teams, Berlin, Germany, (2011).

DÖHLING L. AND LESER U., *EquatorNLP: Pattern-based Information Extraction for Disaster Response,* Terra Cognita 2011 Workshop, Bonn, Germany, (2011).

CARSTEN KRÜGER

Unsicherheiten in der Landnutzungsmodellierung

Ausgangsfragen und Zielsetzung des Projekts:

Zielsetzung 1:
Es sollte ein systematischer Ansatz entwickelt werden, um Unsicherheiten in der Landnutzugsmodellierung zu identifizieren und zu analysieren.

Zielsetzung 2:
Es sollten Methoden entwickelt werden, um Unsicherheiten in Landnutzungs-änderungsprojektionen zu quantifizieren. Dabei sollte zwischen quantitativer Unsicherheit (Wieviel Landnutzung wird sich ändern?) und räumlicher Unsicherheit (Wo wird sich die Änderung ereignen?) unterschieden werden.

Bisher erzielte Ergebnisse:

Zielsetzung 1:
Es wurden Bayesian Belief Netzwerke als vielseitiger Modellierungsansatz identifiziert um die Betrachtung von Unsicherheiten in die Modellierung von Landnutzungsänderung einzubeziehen. Diese Netzwerke beziehen Unsicherheiten in Form von Wahrscheinlichkeiten ein. Sie stellen graphisch die Beziehung zwischen verschiedenen Variablen dar. Die Verbindungen zwischen den Knoten repräsentieren bedingte Wahrscheinlichkeiten. Jede Ausprägung eines Knotens ist abhängig von den Ausprägungen seiner sogenannten Elternknoten. Diese Abhängigkeiten werden über Wahrscheinlichkeitsverteilungen dargestellt. Aus den Wahrscheinlichkeitsverteilungen lassen sich Unsicherheitsmaße wie das Mutual Information Kriterium ermitteln. Mit deren Hilfe wurden bedeute Unsicherheits-quellen der Landnutzungsmodellierung analysiert: die Auswahl der Eingangs-daten, die Ermittlung der Modellstruktur, die Weiterverarbeitung der Ausgangs-daten. Es wurden Wege ermittelt, um die gegenseitigen Beziehungen der

Unsicherheitsquellen untereinander zu bestimmen und ihr Wirken auf das Modellergebnis zu quantifizieren. Das Wirken auf das Modellergebnis wurde dabei unterteilt in quantitative und räumliche Effekte. Der entwickelte Ansatz wurde an Hand eines Fallbeispiels zur Entwaldung des tropischen Regenwaldes im Brasilianischen Amazonas angewendet und getestet.

Zielsetzung 2:
Es wurden Methoden entwickelt, um Unsicherheit in der Modellierung von Landnutzungsänderungen zu quantifizieren. Die Ansätze basieren auf berechneten Wahrscheinlichkeiten und sind anwendbar für zukünftige Zeitpunkte, für die keine Referenzdaten vorliegen. Die Maße unterscheiden zwischen räumlicher und quantitativer Unsicherheit was für die Landnutzungsmodellierung ein enormer Mehrwert ist. Viele Landnutzungsmodelle basieren auf einer quantitativen Nachfragekomponente für das gesamte Untersuchungsgebiet und einem räumlichen Allokationsmechanismus für die geschätzte Nachfrage. Durch die Aufteilung der Gesamtunsicherheit in die beschriebenen zwei Komponenten kann identifiziert werden, welcher Modellteil größere Schwachstellen aufweist.

Die ermittelten Unsicherheitsmaße können aber nicht mit Aussagen über die Modellgenauigkeit gleichgesetzt werden. Ein 100% sicheres Modell kann immer noch falsche Aussagen liefern. Deswegen wurde ein Ansatz entwickelt, um aus der quantifizierten Unsicherheit Aussagen über die Modellgüte ableiten zu können. Es wurde die Beziehung von Unsicherheit und Genauigkeit in bekannten Zeitpunkten genutzt. Für unbekannte zukünftige Zeitpunkte lässt sich nur die Unsicherheit bestimmen. Unter der Annahme, dass der zuvor ermittelte Zusammenhang zeitlich stabil ist, wurde aus den berechneten Unsicherheiten die Modelgenauigkeit für zukünftige Zeitpunkte geschätzt.

Darüber hinaus wurde ein weiterer Ansatz entwickelt, um Unsicherheiten in Landnutzungsprojektionen zu ermitteln. Es wurden Maßzahlen entwickelt um die Wahrscheinlichkeiten zweier Projektionen zu vergleichen. Geringe Unterschiede lassen so auf eine geringe Unsicherheit zukünftiger Entwicklungen schließen, während große Unterschiede auf große Unsicherheiten schließen lassen. Um die ermittelten Unsicherheiten richtig einordnen zu können wurde ein Referenzvergleich eingeführt. Die Referenz basiert auf einem Nullmodell mit einfachen Annahmen, z. B. neue Landnutzungsänderungen ereignen sich dort, wo in der Vergangenheit Landnutzungsänderungen vorzufinden waren. Wenn die Unterschiede zwischen zwei Projektionen ähnlich hoch wie ihre Unterschiede zu einem Nullmodell sind, so kann von einer hohen Unsicherheit ausgegangen werden.

Forschungsaufenthalte, Konferenzen, Workshops:
Zweiwöchiger wissenschaftlicher Austausch an der CLARK University in Worcester (Massachusetts, USA) mit Prof. Robert Gilmore Pontius Jr. Ph.D. zum Thema Unsicherheiten in statistischen Analysen.

Spatial Statistics Conference 2013. Columbus (Ohio), USA: Posterpräsentation, Juni 2013.

6th International Congress on Environmental Modelling and Software (iEMSs). Leipzig, Deutschland: Konferenzartikel, Juli 2012.

Geographic Information Zeitgeist (GIZ), Münster. Deutschland: Konferenzartikel, März 2012.

Eigene Publikationen:

KRÜGER, C., LAKES, T. (accepted for publication). *Revealing Uncertainties in Land Change Modeling using probabilities.* In: Transactions in GIS.

LAKES, T., THEISSELMANN, F.; KÜHNLENZ, F.; KRÜGER, C.; FISCHER, J. (in revision). *Modifying a well-established cellular automata modeling approach to model urban land use dynamics in Greater Tirana,* Albania: exploring benefits of an experiment management. In: Enviromental Modelling and Software.

KRÜGER, C., AND LAKES, T., *Bayesian belief networks as a versatile method for assessing uncertainty in land-change modeling.* International Journal of Geographical Information Science, 29 (1), 111–131, 2014.

KRÜGER, C., *Uncertainty in spatiotemporal modeling of future land- use scenarios*. In: Spatial Statistics Conference 2013. Columbus (Ohio), USA, (2013).

KRÜGER, C., FUNKE, D., LAKES, T., *Approaching uncertainties in land use change modeling in the Amazon rainforest with Bayesian Belief Networks.* In: 6th International Congress on Environmental Modelling and Software (iEMSs). Leipzig, Germany, (2012).

KRÜGER, C., FUNKE, D., LAKES, T., *Modeling Spatio-Temporal Patterns of Deforestation in Brazil with a Bayesian Belief Network Approach*. In: Geographic Information Zeitgeist (GIZ). Münster, Germany, (2012).

MATTHIAS J. SAX

Cloud Based Data Analytics for Disaster Management

In den letzten Jahren hat die Datenstromverarbeitung immer mehr an Bedeutung gewonnen. Unter dem Schlagwort „Big Data“, das mit den 3 Vs Volumen, Velocity und Variety beschrieben wird, bildet Velocity den Datenstromverarbeitungsaspekt mit ab. Im Kontext von METRIK umfasst Datenstromverarbeitung vor allem Sensordaten-Analysen, die sowohl im Katastrophenmanagement als auch in sogenannten „Smart-City“ Anwendungen Bedeutung haben. Die Menge der zu analysierenden Daten ist in den letzten Jahren immer größer geworden, so dass eine Verarbeitung auf einem einzelnen Rechner nicht mehr möglich ist. Deshalb wurden neuartige verteilte Datenstromverarbeitungssysteme entwickelt. Diese Systeme stellen die Forschung vor neue Herausforderungen und Fragestellungen.

Nach dem aktuellen Stand der Technik werden verteilte Datenstromverarbeitungs-systeme manuell aufgesetzt und optimiert. Unter anderem ist dabei die

Clustergröße, also die Anzahl der verwendeten Rechner, ein maßgeblicher Kostenfaktor. Werden zu wenig Rechner eingesetzt, kann der Datenstrom nicht zeitnah verarbeitet werden, was entweder zu hohen Latenzen oder sogar zu (gezielten) Datenverlust führen kann. Gleichzeitig führt eine zu großzügige Bereitstellung von Rechnern zu unnötigen Mehrkosten. Des Weiteren ändern sich die Anforderungen an ein Datenstromverarbeitungssystem über die Zeit (oft zyklisch). Für diesen Fall, ist es wünschenswert die Clustergröße dynamisch anzupassen. Deshalb werden Datenstromverarbeitungssysteme auch oft in einer „Cloud" betrieben, die dynamische Skalierung ermöglicht.

Die große Schwierigkeit besteht nun darin, diesen Prozess zu automatisieren. Dazu ist es notwendig, die benötigte Rechenleistung abzuschätzen und im laufenden Betrieb anzupassen. Diese Kostenabschätzung erweist sich in der Praxis auch für erfahrene Experten als extrem schwierig.

Das Ziel dieser Dissertation ist es, ein Kostenmodell für die Verteilte Datenstrom-verarbeitung zu entwickeln. Dieses Kostenmodell kann dazu verwendet werden, benötigte Rechenkapazitäten abzuschätzen und so den Betrieb von verteilten Datenstromverarbeitungssystemen zu erleichtern. Dabei müssen mehrere Kosten-faktoren wie CPU-Verbrauch, Hauptspeicherverbrauch und Netzwerk Kosten abgeschätzt werden. Gleichzeitig müssen die einzelnen Rechenoperationen effizient auf die zur Verfügung stehenden Ressourcen verteilt werden. Dazu wird im Rahmen der Dissertation auch an Scheduling-Verfahren geforscht. Des weiteren werden Ansätze zum dynamischen Skalieren analysiert und verbesserte Skalierungsalgorithmen entwickelt.

Forschungsaufenthalte, Konferenzen, Workshops:
Forschungsaufenthalt bei Hewlett-Packard Laboratories, Palo Alto (CA), USA, August – Oktober 2013.

Forschungsaufenthalt bei Hewlett-Packard Laboratories, Palo Alto (CA), USA, Mai – Juli 2012.

BTW 2015: 16. Fachtagung Datenbanksysteme für Business, Technologie und Web (BTW), Hamburg, Germany, März 2015.

Poster Präsentation auf dem Adlershofer Forschungsforum, Berlin, Germany, November 2014.

Meeting C++ 2014: 3. Konferenz „Meeting C++", Berlin, Germany, November 2014.

GI FGDB 2013: Herbstreffen der Fachgruppe Datenbanken und Informationssysteme der Gesellschaft für Informatik e. V. (GI), Böblingen, Germany, Dezember 2013.

Stratosphere Summit 2013: 1. „Stratosphere Summit", Berlin, Germany, November 2013.

XLDB 2013: 7th Extremely Large Databases Conference (XLDB), Stanford University, USA, September 2013.

ICDE 2013: IEEE 29th International Conference on Data Engineering (ICDE), Brisbane, Australia, April 2013.

BTW 2013: 15. Fachtagung Datenbanksysteme für Business, Technologie und Web (BTW), Magdeburg, Germany, März 2013.

GI FGDB 2012: Herbstreffen der Fachgruppe Datenbanken und Informationssysteme der Gesellschaft für Informatik e. V. (GI), München, Germany, November 2012.

Stratosphere: Workshop der Stratosphere Forschergruppe, TU Berlin, Germany, November 2012.

GIS Day 2012: GIS day 2012 am Geoforschungszentrum Potsdam (GFZ), Potsdam, November 2012.

VLDB 2012: 38th International Conference on Very Large Databases (VLDB), Istanbul, Turkey, August 2012.

EDBT/ICDT 2012: 15th International Conference on Extending Database Technology and 15th International Conference on Database Theory (joint conference), Berlin, Germany, März 2012.

Stratopshere: Offsite Meeting der Stratosphere Forschergruppe, Döllnsee, Germany, März 2012.

GI FGDB 2011: Herbstreffen der Fachgruppe Datenbanken und Informationssysteme der Gesellschaft für Informatik e. V. (GI), Potsdam, Germany, November 2011.

GDD 2011: Google Developer Day, Berlin, Germany, November 2011.

Stratopshere: Offsite Meeting der Stratosphere Forschergruppe, Neuruppin, Germany, Juli 2011.

VLDB 2011: 37th International Conference on Very Large Databases (VLDB), Seattle, USA, August 2011.

Eigene Publikationen:

MATTHIAS J. SAX, MALU CASTELLANOS: *Building a Transparent Batching Layer for Storm.* Technischer Bericht, Hewlett-Packard Laboratories, HPL-2013-69, Palo Alto (CA), USA, Juli 2014.

ALEXANDER ALEXANDROV, RICO BERGMANN, STEPHAN EWEN, JOHANN-CHRISTOPH FREYTAG, FABIEAN HUESKE, ARVID HEISE, ODE KAO, MARKUS LEICH, ULF LESER, VOKER MARKL, FELIX NAUMANN, MATHIAS PETERS, ASTRID RHEINLÄNDER, MATTHIAS J. SAX, SEBASTIAN SCHELTER, MAREIKE HÖGER, KOSTAS TZOUMAS, DANIEL WARNEKE: *The Stratosphere Platform for Big Data Analytics.* The VLDB Journal, Mai 2014.

MATTHIAS J. SAX, MALU CASTELLANOS, QIMING CHEN, MEICHUN HSU: *Aeolus: An Optimizer for Distributed Intra-Node-Parallel Streaming Systems.* IEEE 29th International Conference on Data Engineering (ICDE), Brisbane, Australia, April 2013.

MATTHIAS J. SAX, MALU CASTELLANOS, QIMING CHEN, MEICHUN HSU: *Performance Optimization for Distributed Intra-Node-Parallel Streaming Systems.* IEEE 29th International Conference on Data Engineering Workshops (ICDEW), Brisbane, Australia, April 2013.

FABIAN HUESKE, MATHIAS PETERS, MATTHIAS J. SAX, ASTRID RHEINLÄNDER, RICO BERGMANN, ALJOSCHA KRETTEK, KOSTAS TZOUMAS: *Opening the Black Boxes in Data Flow Optimization.* Proceedings of the 38th International Conference on Very Large Data Bases, PVLDB 5(11):1256-1267, Istanbul, Turkey, Juli 2012.

CHRISTIAN FIEBRIG, PATRICK KOETHUR, MATTHIAS J. SAX, MIKE SIPS: *Visual Analytics Approach for the Assessment of Simulation Model Output.* Poster at the 3rd International Conference on Data Analysis and Modeling in Earth Sciences (DAMES), Potsdam, Germany, Oktober 2012.

CHRISTIAN BLUM

Selbstorganisation in Netzwerken mobiler Sensorknoten

Ausgangsfragen und Zielsetzung:
Selbstorganisierte drahtlose (ad hoc) Multihopnetzwerke können als einfach einsetzbare, robuste und rekonfigurierbare Kommunikationsinfrastruktur eingesetzt werden, die zum Beispiel in Katastrophenszenarien genutzt werden können wenn die statische Kommunikationsinfrastruktur ausgefallen ist. Die Verbreitung dieser Art von Netzwerken wird zudem durch die fortschreitende Vernetzung von unterschiedlichsten Objekten im Zuge des Internet der Dinge weiter vorangetrieben. Hierbei werden (Gebrauchs-) Gegenstände durch meist drahtlose Netzwerkschnittstellen erweitert was sie in die Lage versetzt drahtlose Netzwerke aufzuspannen. Solche drahtlose Netzwerke nutzen die Luft als geteiltes Kommunikationsmedium was bedeutet, dass die Parameter des Netzwerks sehr orts- und zeit-veränderlich und verrauscht sind. Intelligente Roboter können als Netzwerkknoten diese Herausforderungen meistern indem sie sensomotorische Interaktion ausnutzen indem sie die Sensorinformation aktiv durch ihre Bewegung in der drahtlosen Umgebung formen um dann die Zusammenhänge zwischen Bewegung und Sensorwerten auszunutzen und so die Komplexität zu reduzieren.

Die Problemstellung für diese Arbeit ist daher eine Kontrollstrategie für autonome Roboter zu entwerfen, die die Roboter in die Lage versetzt sich in drahtlosen (ad hoc) Multihopnetzwerken zu integrieren, die nur auf lokaler Information basiert. Diese Kontrollstrategie besteht aus lokalen Kontrollstrategien für jeden einzelnen Roboter und koppelt so die einzelnen Robotern lose durch die sensomotorischen Schleifen. Diese lose Kopplung führt dann zu globaler Selbstorganisation des Systems.

Für einzelne Roboter wurde ein Algorithmus zur Exploration unbekannter Netzwerke für großflächige Außeneinsätze entwickelt und in Simulation evaluiert, wobei sich der Algorithmus als sehr tolerant gegenüber Messrauschen erwiesen hat.

Des Weiteren wurde ein gradientenbasierter Navigationsalgorithmus vor allem für das Finden anderer Netzwerkknoten entwickelt. Die Konvergenzkriterien dieses Algorithmus wurden analytisch bestimmt und er wurde auf einem Roboter implementiert und experimentell in einem Innenraumszenario evaluiert. Es wurde auch gezeigt wie diese Art von Algorithmen sich auf andere Aufgaben als das Finden anderer Netzwerkknoten erweitern lassen können. Zusätzlich wurde ein Metaalgorithmus basierend auf internen Modelle, der Aufgaben, wie zum Beispiel das Finden eines Netzwerkknotens oder die Überbrückung zweier Netzwerk-knoten, lösen kann, entwickelt, implementiert und experimentell in einem Innenraumszenario evaluiert.

Für den Fall von mehreren Roboter, d.h., ein Schwarm drahtlos verbundener Roboter, wurden bereits viele interessante Algorithmen, wie zum Beispiel für die optimale Platzierung mobiler Netzwerkknoten, in der Literatur vorgestellt. Allerdings wird im Moment keiner dieser Algorithmen in realen Szenarien eingesetzt. Einer der Gründe hierfür ist, dass naive Sachwarmalgorithmen meist nicht sicher genau und fehlertolerant sind. Um dieses Problem zu lösen, wurde der Einsatz einer Architektur basierend auf internen Modellen, die einen internen Simulator nutzt, untersucht um die Sicherheit und Fehlertoleranz dieser Systeme zu erhöhen und sie damit geeigneter für reale Anwendung zu machen.

Für diese Architektur wurden zwei Testszenarien untersucht: Im extremen Experiment musste ein Roboter andere Roboter davor bewahren zu Schaden zu kommen und gleichzeitig seine eigene Sicherheit gewährleisten und zusätzlich eine Aufgabe erfüllen. In einem zweiten Experiment musste ein Roboter durch einen engen Korridor navigieren ohne mit anderen Robotern zusammenzustoßen. Die vorgeschlagene Architektur hat sich in diesen Experimenten als sehr effektiv erwiesen.

Forschungsaufenthalte, Konferenzen, Workshops:
Forschungsaufenthalt am Bristol Robotics Laboratory gefördert vom DAAD, Januar bis Juni 2014.

Summer School on Image & Robotics 2011 (SSIR2011), Grenoble, Frankreich, Juli 2011.

DGR-Tage 2011 Deutsche Gesellschaft für Robotik, Karlsruhe, Deutschland, Oktober 2011.

Konferenz zu Unbemannten Autonom-fliegenden Systemen (UAS) für die umwelt-gerechte Landbewirtschaftung (UAS-UL 2012), Berlin, Deutschland, März 2012.

7th German Conference on Robotics (ROBOTIK 2012), München, Deutschland, Mai 2012

DGR-Tage 2012 Deutsche Gesellschaft für Robotik, Berlin, Deutschland, September 2012.

Workshop LC3D „Low-Cost 3D – Sensoren, Algorithmen, Anwendungen“, Berlin, Deutschland, Dezember 2012.

Workshop SACS/SoCoDiS – Conference on Networked Systems (NetSys) – KiVS 2013, Stuttgart, Deutschland, März 2013.

Third EUCogIII Members Conference, Palma de Mallorca, Spanien, April 2013.

Robotics: Science and Systems 2013 (RSS 2013), Berlin, Deutschland, Juni 2013.

9th ACM/IEEE International Conference on Human-Robot Interaction (HRI 2014), Bielefeld, Deutschland, März 2014.

15th Towards Autonomous Robotic Systems (TAROS 2014), Birmingham, United Kindom, September 2014.

Eigene Publikationen:

C. BLUM AND V. V. HAFNER, *An autonomous flying robot for network robotics, in Robotics;* Proceedings of ROBOTIK 2012; 7th German Conference on, pp. 1–5, VDE, 2012.

J. GOSSMANN, C. BLUM, O. BERTHOLD, AND V. V. HAFNER, *Tactile sensors for learning of soft landing on a flying robot,* in Workshop: Sensitive Robotics, Robotics: Science and Systems 2013, 2013.

C. BLUM AND V. V. HAFNER, *Robust exploration strategies for a robot exploring a wireless network,* Electronic Communications of the EASST, vol. 56, 2013.

C. BLUM, O. BERTHOLD, P. RHAN, AND V. V. HAFNER, *Intuitive control of small flying robots,* in Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction, pp. 128–129, ACM, 2014.

C. BLUM AND V. V. HAFNER, *Gradient-based taxis algorithms for network robotics,* arXiv preprint arXiv: 1409.7580, 2014.

F. WINFIELD, C. BLUM, AND W. LIU, *Towards an ethical robot: Internal models, consequences and ethical action selection, in Advances in Autonomous Robotics Systems*, pp. 85–96, Springer, 2014.

MICHAEL FREY

Adaptive Pheromonmodelle in drahtlosen multi-hop Netzwerken

Drahtlose Netzwerke sind eine Schlüsseltechnologie für eine Vielzahl von Anwendungen. Beispiele reichen von Maschennetzwerken zur Breitband-versorgung im ländlichen Kontext über Sensornetzwerke zur Beobachtung von Umweltereignissen oder der Früherkennung von Naturkatastrophen. Zusätzlich vollzieht sich mit dem Internet der Dinge (Internet of Things, IoT) ein Paradigmenwechsel hin zu vernetzten, intelligenten Gegenständen die Teil eines großen Netzwerks werden. Viele jener intelligenten Gegenstände werden drahtlos miteinander kommunizieren. Zu den Herausforderungen des IoTs zählen eine effiziente Kommunikation zwischen heterogenen Geräten unterschiedlichster Ausprägung und eine adaptive und skalierbare Steuerung und Verwaltung von Millionen von Geräten. Hier gilt es neue Ansätze und Verfahren zu entwickeln, die

in der Lage sind die mit dem IoT einhergehendenen Anforderungen zu erfüllen. Von besonderem Interesse sind hierbei Algorithmen die über Eigenschaften der Selbstorganisation verfügen.

Weil viele jener Geräte im IoT energieautark von ihrer Umgebung arbeiten, ist eine energie-effiziente drahtlose Kommunikation unabdingbar. In den letzten Jahren sind dabei eine Vielzahl von verschiedenen Ansätzen im Bereich der drahtlosen Sensornetzwerke entwickelt worden. Im Dissertationsvorhaben werden Strategien entwickelt, die es ermöglichen zur Laufzeit mit Hilfe von Algorithmen auf Basis der Ant-Colony-Optimization- Metaheuristik (ACO-Metaheuristik) die Energieeffizienz in drahtlosen Netzwerken zu verbessern. Die Strategien bestehen aus zwei Komponenten. Zum einen ein Modell, dass es ermöglicht auf Basis des aktuellen Datenverkehrs vorherzusagen, wie der Zustand einzelner benachbarter Knoten im Netzwerk ist und zum anderen aus einer Komponente, die es erlaubt, auf Basis jener Modellinformation konkrete Anpassungen an Parametern des Algorithmus vorzunehmen. Vorbild für die Metaheuristik ist das Futterverhalten von Ameisen, die auf ihrem Weg von der Ameisenkolonie zu einer Nahrungsquelle den Weg mit einem speziellen Hormon, genannt Pheromon, markieren. Andere Ameisen der Kolonie werden von Pheromonen angelockt und verstärken wiederum auf ihren Weg zur Nahrungsquelle die Pheromonspur. Allerdings verdunsten Pheromone und über einen Zeitraum hinweg bildet sich damit der kürzeste Pfad von der Ameisenkolonie zur Nahrungsquelle. ACO-basierte Algorithmen zählen zur Klasse der Schwarmintelligenzalgorithmen und erfüllen die Anforderungen an ein selbstorganisierendes System.

Kernbestandteil des Dissertationsvorhabens ist die Entwicklung von Pheromon-modelle für drahtlose multi-hop Netzwerke. In der Natur ermöglichen Pheromone eine indirekte Kommunikation (Stigmergie) der Ameisen über die Umgebung. Übertragen auf drahtlose multi-hop Netzwerke ergibt sich die Möglichkeit anhand der Entwicklung der Pheromone lokal, Zustände im Netzwerk herzuleiten. Die Konzentration von Pheromonen auf verschiedenen Pfaden, aber auch die Verstärkungs- und Verdunstungsprozesse, haben konkrete Auswirkungen auf das Verhalten ACO-basierter Algorithmen. Zusätzlich beeinflusst aber auch die Art wie und welche Werte den Pheromonen in initialen Phasen zugeordnet werden die Performanz des Algorithmus. Pheromonmodelle sind im Gegensatz zur Biologie in der Informatik nicht eindeutig spezifiziert. Im Rahmen des Dissertationsvorhabens wird ein Pheromonmodell als ein analytisches Modell definiert, das die Entwicklung von Pheromonen über einen definierten Zeitraum beschreibt. Pheromonmodelle ermöglichen es, Aussagen darüber zu treffen, wie sich Ameisen einer Ameisenkolonie in ihrer Futtersuche verhalten. Schwerpunkte der Arbeit bilden die Fragen inwieweit sich Pheromonmodelle eignen, um die Energieeffizienz in drahtlosen multi-hop Netzwerken zu verbessern und Zustände im Netzwerk zu

erkennen. Weiterhin wird analysiert, wie sich Steuerungsmechanismen nutzen lassen, um ACO-basierte Routing-Algorithmen adaptiv zu regeln. Mit den Verstärkungs- und Verdunstungsprozessen existieren bereits implizit Feedbackmechanismen zur Steuerung von Ameisenalgorithmen, allerdings verändern sich die Parameter jener Prozesse zur Laufzeit nicht.

Exemplarisch werden im Dissertationsvorhaben Pheromonmodelle anhand des Ant Routing Algorithm (ARA) und seiner Erweiterung, dem Energy-Aware Ant Routing Algorithm (EARA) [Frey2014a] untersucht. Letzterer berücksichtigt bei der Weiterleitung von Paketen nicht nur Pheromone, sondern auch die verbleibende Energie benachbarter Netzwerkknoten. Mit libARA [Frey2013b] wurde ein Framework zur Untersuchung von ACO-basierten Routing- Algorithmen entwickelt. Das Framework erlaubt erstmalig eine Untersuchung der Algorithmen sowohl in Simulation als auch in einer Testbedumgebung. Das Framework stellt dabei modulare Komponenten bereit, um die grundlegenden Konzepten von ACO-basierten Routing-Algorithmen abzubilden. Die Implementierung von weiteren ACO-basierten Routing-Algorithmen für Simulation und Testbed wird damit erheblich erleichtert.

Für Untersuchungen in experimentellen drahtlosen Testumgebungen wird das DES-Testbed an der Freien Universität Berlin genutzt. In einer ersten Voruntersuchung konnte gezeigt werden, dass die sehr guten Ergebnisse von ARA in simulationsbasierter Studien sich in ähnlichen Szenarien in einer Testbedumgebung nicht reproduzieren lassen. Auf Basis der Messungen wurde eine adaptive Pfadklassifikation eingeführt und alternative Verdunstungsfunktionen spezifiziert. Die adaptive Pfadklassifikation unterteilt dabei Pfade in effiziente, akzeptable und verlustbehaftete Pfade. Für effiziente Pfade wird die Verdunstung gebremst, wohingegen für verlustbehaftete Pfade die Verdunstung der Pheromone beschleunigt wird. Mit Hilfe der adaptiven Pfadklassifikation konnte eine Leistungsverbesserung erzielt werden, allerdings nicht in ausreichendem Umfang.

Aufgrund der Natur des DES-Testbeds wurde der Ansatz bisher nur in drahtlosen Maschennetzen untersucht. Studien zur Energieeffizienz ACO-basierter Algorithmen sind aufgrund der eingesetzten Hardware im DES-Testbed nur eingeschränkt möglich. Ziel des Forschungsaufenthaltes am SICS Swedish ICT war es daher, das Forschungsvorhaben auf drahtlose Sensornetze zu übertragen und die Energieeffizienz in einer drahtlosen experimentellen Sensornetzumgebung zu untersuchen. Während des Aufenthaltes am SICS Swedish ICT wurde daher das libARA Framework für Betriebsystem für drahtlose Sensornetzwerke portiert. Eine erste Evaluation der Portierung des Ansatzes für drahtlose Sensornetze konnte noch während des Aufenthalts am SICS Swedish ICT durchgeführt werden. Weiterhin

entstand unter anderem eine Publikation [Frey2015b] zur Leistungsfähigkeit von benachbarten IEEE 802.15.4 Netzwerken.

Für die Dissertation müssen einige Experimente wiederholt oder erweitert werden. Einzelne Kapitel der Dissertationsschrift wurden bereits begonnen und die Fertigstellung ist innerhalb eines Jahres geplant. Das Dissertationsvorhaben ist mit keinen anderen Projekten des Graduiertenkollegs vernetzt.

Forschungsaufenthalte, Konferenzen, Workshops:
Aufenthalt am SICS Swedish ICT in der Networked Embedded Systems Group, Stockholm, Schweden (DAAD FITweltweit Stipendium), Juli – Dezember 2014.

IEEE International Conference on Communications (ICC), Sydney, Australien, Juni 2014.

OMNeT++ Community Summit, Hamburg, September 2014.

International Conference on Performance Evaluation Methodologies and Tools (ValueTools), Turin, Italien, Dezember 2013.

IEEE International Conference on Computer Communications (INFOCOM), Turin, Italien, April 2013.

ACM Conference on Embedded Networked Sensor Systems (SenSys), Seattle, USA, November 2011.

Jahrestagung der Gesellschaft für Informatik, Berlin, Oktober 2011.

International Summer School Cooperating Objects Network of Excellence International Summer School "Networked Embedded Systems: Humans in the Loop", Bertinoro, Italien, Juli 2011.

Eigene Publikationen:
LAURA M. FEENEY, MICHAEL FREY, VIKTORIA FODOR, MESUT GÜNES: *Modes of Inter-network Interaction in Beacon-Enabled IEEE 802.15.4 Networks,* Proceedings of the 14th IFIP Annual Mediterranean Ad Hoc Networking Workshop (Med-Hoc-Net 2015), Juni 2015.

MICHAEL FREY, MESUT GÜNES: *Follow the Pheromone Trail – On Studying Ant Routing Algorithms in Simulation and Wireless Testbeds,* Proceedings of the 18th Communications & Networking Symposium, part of the 2015 Spring Simulation Multiconference, SpringSim '15, Alexandria, VA, USA, April 2015.

MICHAEL FREY, MESUT GÜNES, FRIEDRICH GROSSE: *Towards Energy-Aware Ant Routing in Wireless Multi-Hop Networks,* Technischer Bericht, Universität Münster, Dezember 2014.

MICHAEL FREY, MESUT GÜNES: *Attack of the Ants: Studying Ant Routing Algorithms in Simulation and Wireless Testbeds,* Proceedings of the First OMNeT++ Community Summit 2014, September 2014.

MICHAEL FREY, FRIEDRICH GROSSE, MESUT GÜNES: *Energy-Aware Ant Routing in Wireless Multi-Hop Networks,* Proceedings of the IEEE International Conference on Communications (ICC), Juni 2014.

MICHAEL FREY, FRIEDRICH GROSSE, MESUT GÜNES: libARA: *A framework for simulation and testbed based studies on ant routing algorithms in wireless multi-hop networks, Proceedings of the 7th International Conference on Performance Evaluation Methodologies and Tools,* Dezember 2013.

MICHAEL FREY, MESUT GÜNES: *Combining Control Loops and Ant Algorithms to Optimize Network Lifetime,* Proceedings of the 32nd IEEE International Conference on Computer Communications (INFOCOM), April 2013.

ANDREAS TEXTOR, FABIAN MEYER, MICHAEL FREY, MARCUS THOSS, JAN SCHAEFER, REINHOLD KROEGER: *An Architecture for Semantically Enriched Data Stream Mining,* Proceedings of the 1st International Conference on Data Analytics, September 2012.

MICHAEL FREY: *Self-Adaptive Wireless Sensor Networks,* Doctoral Colloqium at the 9th Conference on Embedded Network Sensor Systems (SenSys 2011), November 2011.

MARTIN R. KOST

Privacy aware Data Processing - Analysis and Implementation of Privacy Protection using Ontologies

In Zukunft bestehen intelligente Transport Systeme (ITS) aus Fahrzeugen, Straßenrandeinheiten (engl. roadside units), Netzwerkknoten und serverseitigen Diensten. Teilnehmer eines ITS-Systems tauschen Informationen miteinander aus, um erweiterte Funktionalitäten (wie z. B. verbesserte Verkehrsdienste, Fahrassistenten, Transportoptimierung) anbieten zu können. Dienste, welche erweiterte Funktionalitäten realisieren, nehmen Einfluss auf die Privatsphäre von Personen, da sie personenbezogene Informationen (wie z. B. Ortsinformationen) verarbeiten. Dabei begünstigt eine nicht-regulierte Informationsverarbeitung mögliche Verletzungen der Privatsphäre. Aus diesem Grund wird der Schutz der Privatsphäre durch die europäische Rechtsprechung adressiert und behandelt [European Parliament and Council. Directive 2010/40/EU. Official Journal L 207]. Das Ziel der Arbeit ist es, die Umsetzung der Gesetzgebung zum Schutz der Privatsphäre durch technische Maßnahmen zu vervollständigen.

Existierende Lösungen zum Schutz der Privatsphäre schützen lediglich das Ereignis des Datenzugriffs und berücksichtigen nicht individuelle Anforderungen an den Schutz der Privatsphäre für den gesamten Datenfluss in (verteilten) Systemen. Im Gegensatz dazu wird in der Doktorarbeit die Kontrolle des gesamten Datenflusses realisiert. D. h. es werden zusätzlich Ereignisse, wie das Kommunizieren und Prozessieren von Daten, durch unterschiedliche Anwendungen oder entfernte Knoten in die Kontrolle eingebunden. Zur Umsetzung dieser erweiterten

Kontrollmechanismen führen wir ein neues formales Modell für die Anfrageverarbeitung ein. Mit diesem Modell integrieren wir Aspekte für den Schutz der Privatsphäre direkt in die Anfrageverarbeitung. Zunächst haben wir u.a. Ansätze zur Formalisierung unseres Modells untersucht und ausgewählt. Die meisten existierenden Ansätze zum Schutz der Privatsphäre basieren auf einer zentralisierten Zugriffskontrolle oder besitzen keine Formalisierung. Verwandte Bereiche wie Model Checking oder Programmiersprachen bieten Techniken, mit denen wir eine große Zahl an Systemeigenschaften verifizieren können. Allerdings besitzen diese Techniken eine hohe Komplexität und betrachten nicht direkt die Auswertung von technischen Ansätzen zum Schutz der Privatsphäre, wie z. B. das Durchsetzen/Erzwingen von individuellen Schutzkriterien.

In der Doktorarbeit werden neue Ansätze eingeführt, welche technische Details zur formalen Analyse von ITS-Systemen auswerten. Zur Beschreibung und Abstraktion werden dabei modellbasierte und semantische Technologien (z.|B. in Form von Komponenten- und Datenflussmodellen sowie Ontologien) eingesetzt. Bei der Anwendung dieser Ansätze werden System-Modelle verarbeitet, die Informationen über das System wie z. B. seine Datenflüsse, Kompositionen und verarbeitete Datentypen enthalten. Unter Verwendung einer Auswahl an Modellelementen werden domänenabhängige und domänenunabhängige Bedingungen/Kriterien und Indikatoren definiert und ausgewertet. Diese Bedingungen und Indikatoren können von Prinzipien zum Schutzes der Privatsphäre durch Anwendung von speziellen Entwurfsprozessen (“Privacy by Design”) abgeleitet werden. Für ein gegebenes System-Modell wird die Einhaltung geforderter Bedingungen verifiziert.

Zur Umsetzung von erweiterten Kontrollmechanismen wird zudem ein Ansatz eingeführt, welcher Aspekte für den Schutz der Privatsphäre in die Anfrageverarbeitung integriert. Die Arbeit beinhaltet dazu ein einfaches Anfrageverarbeitungsmodell, welches die Privatsphäre während der Abarbeitung einer Anfrage im Einzelbenutzermodus schützt. Anstatt Privatsphäre als Zusatz oder Ergänzung zu einem System zu behandeln, sind Aspekte zum Schutz der Privatsphäre direkt als Basiskonzepte (z.B. individuelle Schutzkriterien) in das Modell integriert. Das Ziel bei der Verwendung des Modells ist es 1.) eine gegebene Anfrage bezüglich ihrer Auswirkungen auf die Privatsphäre zu untersuchen; z.|B. wird überprüft, ob die Anfrage gegebene individuelle Schutzkriterien einhält und es werden die Kosten sowie Risiken für die Privatsphäre berechnet, welche für die Ausführung der Anfrage erwartet werden; und 2.) eine gegebene Anfrage so umzuschreiben bzw. durch zusätzliche Operationen (privacy operators) zu erweitern, dass die gegebenen individuellen Schutzkriterien erzwungen/umgesetzt werden; z.|B. werden Anonymisierungsoperationen eingefügt, um Schutzkriterien wie k-Anonymität umzusetzen. Für diesen Zweck wird eine gegebene Anfrage durch einen entsprechen-den algebraischen Ausdruck formalisiert. Solch ein

Ausdruck verwendet Standardoperatoren und Konstrukte der relationalen Algebra, welche um Aspekte zum Schutz der Privatsphäre erweitert sind. Diese Aspekte können beispielsweise Bedingungen bzgl. der Attribute definieren, welche ein join-Operator einliest oder produziert. Mithilfe solcher Bedingungen soll verhindert werden, dass eine Anfrage Daten auf eine Weise kombiniert, welche die Privatsphäre verletzt; z. B. indem durch die Ausführung der Anfrage sensitive Informationen über eine Person identifiziert werden. Für das Erzeugen von Ausdrücken, die konform zu den erhobenen Anforderungen sind, erweitern wir die relationale Algebra um entsprechende Bedingungen und Regeln. Ergänzend zum beschriebenen (passiven Schutz durch) Definieren und Auswerten von Bedingungen, entwickeln wir Mechanismen, welche aktiv die Umsetzung der Anforderungen zum Schutz der Privatsphäre unterstützen. Dazu werden die erzeugten Ausdrücke adaptiert; z.|B. durch das Einfügen von Anonymisierungs- oder Verschlüsselungsoperatoren. Wesentlicher Bestandteil für die Auswertung und die zielgerichtete Adaption/Umschreiben von Anfragen sind die veränderten Definitionen von Korrektheit einer Anfrage bzw. von semantischer Äquivalenz zweier Anfragen.

Die prototypische Umsetzung dieses Ansatzes umfasst die Verwendung der Architektur PeRA (Privacy-enforcing Runtime Architecture). Die PeRA-Architektur setzt den Schutz der Privatsphäre zur Laufzeit durch indem sie einen Perimeter schafft in welchem die Einhaltung von Policies garantiert wird. Policies definieren dabei Kriterien zum Schutz der Privatsphäre. Personen die dem System Daten über sich zur Verfügung stellen erhalten somit die Kontrolle darüber wie Anwendungen ihre Daten verarbeiten können. Personenbezogene Daten werden innerhalb des Perimeters untrennbar mit den - durch die zugehörigen Personen definierten - Policies verbunden. Kernkomponenten der PeRA- Architektur stellen sicher, dass Anwendungen nur in einer Policy-kompatiblen Art Operationen auf Daten ausführen können. Eine dieser Kernkomponenten ist die Anfrage- und Policy-Analyse-Einheit, welche Aussagen einer deklarativen Anfragesprache zusammen mit Aussagen von Policysprachen verarbeitet. Bei der Auswertung ob eine Anfrage ausgeführt werden darf oder nicht wird der Anfragekontext mit einbezogen. Dieser wird durch Metadaten beschrieben, welche vom System zur Verfügung gestellt und verifiziert werden.

Forschungsaufenthalte, Konferenzen, Workshops:
Poster Präsentation auf dem Adlershofer Forschungsforum, Berlin, Deutschland, November 2013.

BTW 2013: 15. Fachtagung Datenbanksysteme für Business, Technologie und Web (BTW), Magdeburg, Deutschland, März 2013.

EDBT/ICDT 2012: 15th International Conference on Extending Database Technology and 15th International Conference on Database Theory (joint conference), Berlin, Deutschland, März 2012.

CODASPY-Konferenz San Antonio, USA, Februar 2012.

GDD 2011: Google Developer Day, Berlin, Deutschland, November 2011.

GI-Jahrestagung Berlin, Deutschland, Oktober 2011.

ARES-Konferenz Wien, Österreich, August 2011.

Eigene Publikationen:

S. Dietzel, M. Kost, F. Schaub, and F. Kargl. CANE: A Controlled Application Environment for Privacy Protection in ITS. In 12th International Conference on ITS Telecommunications (ITST 2012). IEEE, 2012.

D. Janusz, M. Kost, and J.-C. Freytag. Privacy protocol for linking distributed medical data. In W. Jonker and M. Petkovic, editors, Secure Data Management, volume 7482 of Lecture Notes in Computer Science, pages 45–57. Springer, 2012.

M. Kost, R. Dzikowski, and J.-C. Freytag. PeRA: Individual privacy control in intelligent transportation systems. In Proceedings of the Demonstration Session at the 15. GI-Fachtagung Datenbanksysteme für Business, Technologie und Web (BTW), 2013.

M. Kost and J. C. Freytag. Privacy analysis using ontologies. In Proceedings of the second ACM conference on Data and Application Security and Privacy, CODASPY '12, pages 205– 216, New York, NY, USA, 2012. ACM.

M. Kost, J.-C. Freytag, F. Kargl, and A. Kung. Privacy Verification using Ontologies. In Proceedings of the First International Workshop on Privacy by Design, Vienna, Austria, 2011.

M. Kost, B. Wiedersheim, S. Dietzel, F. Schaub, and T. Bachmor. PRECIOSA PeRA: Practical enforcement of privacy policies in intelligent transportation systems. In Proc. of the Demo. Session at the Fourth ACM Conf. on Wireless Network Security, 2011.

VLADICA SARK

Localization in Wireless Sensor Networks Using Time of Fligth Measurements

Für ein Ortungssystem auf Basis von Laufzeitmessungen gibt es noch weitere zu lösende Aufgaben. Erstens führt die endliche Abtastrate zu einer hohen Granularität im gemessenen Abstand (range binning). Zweitens verursachen Ungenauigkeit und Drift in der Systemuhr weitere Messfehler. Zwar kann dieser Effekt theoretisch durch hochgenaue Clock-Generatoren vermieden werden, aber Lösungen wie Atomuhren kommen bei portablen Geräten und Sensorknoten nicht in Frage. Da Fehler durch die ungenaue Systemuhr prinzipiell nicht komplett

vermieden werden können, ist es wichtig, diese Fehler abzuschätzen und zu minimieren. Bei Methoden wie TWR ist dies eine signifikante Herausforderung. Schließlich wird das System durch Faktoren in der Umgebung und im Übertragungskanal limitiert. Genau genommen ist dies das angesprochene Problem der Nichtsichtverbindungen (NLOS). Wenn der Abstand zwischen zwei Geräten bestimmt werden soll und es keine Sichtverbindung gibt, so wird das Resultat zu höheren Werten verschoben. Die Ursache ist, dass nur reflektierte Radiowellen bei dem Empfänger einlaufen. Eine direkte Bestimmung des tatsächlichen Abstandes ist dann nicht ohne weiteres möglich.

In dieser Dissertation werden hauptsächlich die Laufzeitmethoden (TOF) zur Ortung und Abstandsmessung untersucht. Das Hauptziel ist, eine höhere Genauigkeit zu erreichen als bei alternativen Methoden, welche die Abnahme der Wellenintensität auswerten. Um Synchronisationsproblemen aus dem Weg zu gehen, liegt der Fokus auf TWR-Verfahren, da bei diesen keine präzise Zeitsynchronisation zwischen den Knoten erforderlich ist. Das grundlegende Verfahren ist, die Abstände zu mehreren Ankerpunkten mit bekannten Positionen zu bestimmen. Eine nachfolgende Lateration führt zu einer Bestimmung der Position des Sensorknotens.

Eine Pseudozufalls- (engl. pseudo-random, PR) Sequenz wird zur Abstandsbestimmung eingesetzt. Dieses Verfahren ist ähnlich den Frequenzspreizverfahren in der Datenübertragung (z.|B. direct sequence spread spectrum, DSSS). Im vorliegenden Fall werden keine Daten aufmoduliert. Die PR Sequenz wird von dem Knoten ausgestrahlt, welcher die Abstandsmessung initiiert. Der zweite Knoten empfängt die Sequenz und sendet sie zurück an den initiierenden Knoten. Im Allgemeinen wird BPSK (binary phase-shift keying) oder QPSK (quadrature phase-shift keying) zur Übertragung der PR Sequenz verwendet.

Anderseits gibt es bereits einfachere Alternativmethoden, welche recht gute Resultate erreichen können. Das Grundprinzip dieser Methoden wird als Dauerstrichradar bezeichnet (continuous wave, CW, oder dual-frequency continuous wave, DFCW, radar). Im Prinzip sind dies TWR Methoden, welche Sinuswellen an Stelle von PR Sequenzen einsetzen. Dadurch erfordern sie keine hohe Rechenleistung und können mit kostengünstiger Hardware verwirklicht werden. Der Hauptnachteil ist, dass sie bei einer Mehrwegeausbreitung nur schlecht funktionieren.

Die vorliegende Dissertation behandelt die angesprochenen Probleme bei TWR Methoden, welche PR Sequenzen verwenden. Erstens wird die unerwünschte Granularität in der Abstandsmessung (range binning) aufgrund der endlichen Abtastrate verbessert. Im System wird das Signal von digital zu analog und umgekehrt konvertiert. Wenn die digital-zu-analog und analog-zu-digital Konverter eine zu niedrige Abtastrate verwenden, können nur grobe

Abstandsmessungen durchgeführt werden. Die Standardlösung dieses Problems setzt eine Interpolation ein. Um diese Interpolation mit ihrem zugehörigen Rechenaufwand zu umgehen, wurde ein neues Verfahren namens „Modified Equivalent Time Sampling“ entwickelt [Sark2013]. Die Methode benötigt zwar eine längere Zeit, um eine Messung durch zu führen, reduziert aber den Aufwand in der Verarbeitung des empfangenen Signals. Dadurch wird der Ablauf einer Messung soweit vereinfacht, dass eine Implementation auf drahtlosen Sensorknoten mit beschränkten Ressourcen ermöglicht wird.

Zweitens wird die Integration der Abstandsmessung in ein System zur drahtlosen Datenübertragung untersucht. In vielen bestehenden Verfahren wird ein Datenpaket ausgesendet, welches dann von dem zweiten Knoten mit einem Antwortpaket (acknowledgement, ACK) bestätigt wird. Das Zeitintervall zwischen dem Aussenden des ersten Paketes bis zum Empfang der Bestätigung wird gemessen und daraus die Laufzeit zwischen den Knoten bestimmt. Dieses Verfahren ist zwar einfach, aber nicht optimal, da ein Standard-Datenpaket aus Sicht der Abstandsmessungen nicht die beste Wahl ist. Andere Verfahren basieren auf DSSS für die Datenübertragung. Die Nutzdaten modulieren die PR Sequenz und das Resultat wird ausgestrahlt. Die PR Sequenz wird gleichzeitig verwendet, um die Abstandsmessung wie besprochen durch zu führen. Dieses Verfahren liefert bessere Resultate, kann aber bisher nur in DSSS Systemen verwendet werden. Andere Systeme, welche nicht auf DSSS basieren, können diesen Zugang nicht implementieren. In der vorliegenden Arbeit wird ein neues Verfahren vorgeschlagen, welches bei einer Datenübertragung im 60 GHz Band die Vorteile einer präzisen Lokalisierung und einer hohen Datenrate verbindet. Die Datenübertragung verwendet ein orthogonales Frequenzmultiplexverfahren (orthogonal frequency-dicision mutliplexing, OFDM). Die Grundidee ist, dass getrennte Zeitintervalle (time slots) für die Datenübertragung und für die Abstandsmessung genutzt werden. Die beiden beteiligten Knoten einigen sich darüber, zu welchen Zeiten Daten übertragen werden und wann eine Abstandsmessung durchgeführt wird. Dieser Zugang erlaubt, unabhängige Modulationsverfahren für die beiden Teilaufgaben zu verwenden. Als weiterer Vorteil sind Entwicklung und Implementierung der Systeme zur Abstandmessung und zur Datenübertragung weitgehend unabhängig von einander. Das Verfahren wurde in einem 60 GHz Demonstrator verwirklicht [Sark2014b]. Die Abstandsmessung erreicht eine Genauigkeit von etwa 1 cm bei einer Entfernung bis zu 10 m. Die vergleichsweise große Bandbreite von 2 GHz ist der Hauptgrund für das Erreichen einer hohen Präzision.

Ein verbreitetes Problem bei der Simulation von RF (radio frequency) Abstandsmessungen ist das verwendete Kanalmodell. Die verfügbaren Modelle wurden in erster Linie für die drahtlose Datenkommunikation entwickelt. Sie sind für ToF

Systeme weniger geeignet. Zwar können ToF Messungen simuliert werden, aber die Resultate entsprechen nicht der Wirklichkeit. Ein weiteres Problem stellt sich bei der Verifikation der Algorithmen zur Abstandsmessung in realistischen Szenarien. Im Allgemeinen werden solche Algorithmen in programmierbarer Hardware implementiert, etwa in einem Field Programmable Gate Array (FPGA) [Sark2014c]. Diese Prozedur ist kompliziert und zeitaufwendig. Um effektiv die Leistungsfähigkeit der verschiedenen Algorithmusvarianten unter realen Bedingungen zu testen, wurde deshalb in dieser Arbeit ein Software Defined Radio (SDR) eingesetzt [Sark2014a],[Sark2015a]. Dieses SDR wurde modifiziert, um Abstandsmessungen zu ermöglichen. Im Laufe der Arbeit wurde dann weitere Funktionalität hinzugefügt, welche eine effiziente Implementierung von Ortungsalgorithmen unterstützt. Hauptziel war dabei, die aufwendige FPGA-Programmierung bei der Untersuchung eines neuen oder abgeänderten Algorithmus zu umgehen.

Als letztes Thema dieser Dissertation wurden kooperative Ortungsverfahren (cooperative localization) untersucht. Damit bezeichnet man Verfahren, bei welchen alle Knoten in einem drahtlosen Sensornetz kooperieren, um ihre relativen Positionen zu ermitteln. Die Hauptschwierigkeit in solchen Methoden sind die Abweichungen der Systemuhren (Clock) zwischen den Knoten, welche Fehler in den ermittelten Abständen zur Folge haben. Es ist ein Merkmal der kooperativen Verfahren dass diese Fehler propagieren und mit steigender Knotenzahl zunehmen. Mittels Simulation wurde diese Frage untersucht. Die Resultate zeigen, dass für mehr als fünf Knoten und bei Verwendung von kommerziell erhältlichen Quarz-Oszillatoren die Fehler so weit anwachsen, dass die Positionsbestimmung unbrauchbar wird. Zur Lösung des Problems wird eine neue Methode zur Reduzierung der Fehler vorgeschlagen. Hierbei werden die Zeitverschiebungen zwischen den Systemuhren verwendet, um die Positionsfehler zu korrigieren. Die Zeitverschiebung zwischen zwei Knoten wird normalerweise berechnet, wenn ein Datenpaket empfangen wird. Diese Information spielt bei der Datenkommunikation eine Rolle, kann aber darüber hinaus zur Kompensation der Lokalisierungsfehler verwendet werden. Simulationen haben nachgewiesen, dass mit diesem Verfahren der Positionsfehler zu akzeptablen Werten verkleinert werden. Als weiterer Vorteil wird die Anzahl der Datenübertragungen durch die Methode nicht erhöht, wodurch Energie und Bandbreite eingespart wird.

Zukünftige Untersuchungen im Bereich der Lokalisierung wären besonders sinnvoll im Zusammenhang mit CW und DFCW Verfahren. Diese sind einfach zu implementieren, sind aber in Umgebungen mit hoher Mehrstrahlinterferenz nicht verlässlich. Eine Kombination der CW/DFCW Methoden mit dem Einsatz von PR Sequenzen wie bei DSSS sollte die Abstandsmessung robust gegen Mehrstrahlinterferenz machen, während gleichzeitig die Genauigkeit erhöht wird.

Schließlich ist das Thema Lokalisierung von hoher Relevanz für das Graduiertenkolleg METRIK. Eine präzise Ortung und Abstandsmessung ist von Vorteil in vielen Katastrophenszenarien.

Forschungsaufenthalte, Konferenzen, Workshops:
International Symposium on Signals, Systems and Electronics, Gran Canaria, Spain, Mai 2015.

Microsoft Indoor Localization Competition, The 13th ACM/IEEE International Conference on Information Processing in Sensor Networks, Berlin, April 2014.

24th Annual IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC 2013), London, UK, September 2013.

„KMU und Forschung: Gemeinsam zu intelligenten Lösungen“, Fachtagung KMU-innovativ: IKT 2014, November 2014.

Eigene Publikationen:
VLADICA SARK, ECKHARD GRASS, *Configurable Software Defined Radio System for two Way Time of Flight Ranging,* Proceedings of the International Symposium on Signals, Systems and Electronics (ISSSE 2015) abstr., Gran Canaria, Spain, May 2015.

VLADICA SARK, ECKHARD GRASS, *A Software Defined Radio for Time of Flight Based Ranging and Localization,* Proceedings of the Microsoft Indoor Localization Competition - IPSN 2014 abstr., Berlin, Germany, April 2014.

M. EHRIG, V. SARK, J. GUTIERREZ TERAN, M. PETRI, E. GRASS, *A 60 GHz System for Simultanious Time of Flight Ranging and High-Speed Wireless Data Communication,* Proceedings of the Microsoft Indoor Localization Competition - IPSN 2014 abstr., Berlin, Germany, April 2014.

M. ERIGH, M. PETRI, V. SARK, J. G. TERAN, E. GRASS, *Combined high-resolution ranging and high data rate wireless communication system in the 60 GHz band,* Proceedings of the 11th Workshop on Positioning, Navigation and Communication 2014 (WPNC’14), Dresden, Germany, 12-13 March 2014.

M. METHFESSEL, V. SARK, G. FISCHER, *Empfänger, Anordnung und Verfahren für die Ultrabreitband-Übertragung,* IHP.384.PCT-Anmeldung am 29.04.2014, AZ: PCT / EP2014 / 058739.

VLADICA SARK, ECKHARD GRASS, *Modified Equivalent Time Sampling for Improving Precision of Time-of-Flight Based Localization,* Proceedings of the IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC 2013), London, UK, September 2013.

Z. STAMENKOVIC, K. TITTELBACH-HELMRICH, J. DOMKE, C. LÖRCHNER-GERDAUS, J. ANDERS, V. SARK, M. ERIC, N. SIRA, *Rear View Camera System for Car Driving Assistance,* Proceedings of the 28th International Conference on Microelectronics (MIEL 2012), Nis, Serbia, May 2012.

OSWALD BERTHOLD

Robotic Self-Exploration and Acquisition of Sensorimotor Primitives

Ausgangsfrage und Zielsetzung

Die Funktion eines Roboters und seines Verhaltens ist, unabhängig von einer konkreten Aufgabe, untrennbar mit seinen sensomotorischen Fähigkeiten verknüpft. Diese ermöglichen es einem Roboter, Verhalten mittels koordinierter Motorsignale zu erzeugen, die wiederrum von der aus der Umwelt aufgenommenen Information beeinfluss und geformt werden. Da die explizite Programmierung dieser Fähigkeiten nicht zielführend ist, muss ein Roboter in der Lage sein, sie sich selbst anzueignen und durch kontinuierliche Selbst-Exploration zu verbessern.

Das Problem lässt sich wie folgt formulieren: Ich betrachte ein Robotersystem mit gegebenen Sensoren, Motoren und inhärenter Dynamik und suche ein internes inverses Modell (Abbildung oder System), das den sensorischen Input auf Motorsignale abbildet, sodass eine gegebene, im Sensorraum definierte Kostenfunktion minimiert wird. In Unterscheidung zum klassischen Reinforcement Learning, betrachte ich zum einen Modelle mit Gedächtnis (Systeme) und zum anderen suche ich nicht notwendigerweise ein Minimum sondern Punkte mit Kosten unterhalb eines Schwellwerts.

Ansatz

Ich kombiniere Methoden aus der Entwicklungsrobotik (Developmental Robotics, DR) mit denen des Verstärkungslernens (Reinforcement Learning) und der Merkmalsextraktion (Feature Learning). Im Bereich DR bildet die Methode der internen Modelle ein allgemeines Konzept zur Behandlung und Synthese von Verhalten. Das heißt, ein Agent erzeugt eines oder mehrere innere Modelle der Abhängigkeiten zwischen Motor- und Sensorsignalen und nutzt dann diese Modelle um selbstgestellte oder von außen gesetzte sensorische Ziele zu erreichen. In meiner Arbeit bezeichne ich ein solches Modell zusammen mit seinem Kontext als sensomotorische Primitive. In meinem Ansatz werden Online-Merkmals-extraktion und Verhaltenslernen in einer geschlossenen Regelschleife inhärent kombiniert. Der Ansatz ist sowohl von neurowissenschaftlichen Prinzipien als auch von Beobachtungen tierischer Entwicklungsprozesse motiviert. Zur Erzeugung komplexer Verhalten können mehrere Primitiven gleichzeitig aktiv sein und hierarchisch organisiert werden.

Resultate

Der Ansatz wurde bereits erfolgreich auf verschiedenen Systemen angewendet. Diese beinhalten die Bewegungsregelung eines Flugroboters mittels einer mitfliegenden Kamera, die unüberwachte Aneignung einer Kamerabelichtungs-regelung, das unüberwachte Erlernen eines Eigenbewegungsschätzers auf Basis

dichter optischer Flussfelder und die Aneignung von Stabilisierungsverhalten auf simulierten und echten Robotern.

Forschungsaufenthalte, Konferenzen, Workshops:
Jahrestagung der Deutschen Gesellschaft für Robotik in Berlin, Oktober 2012.

Symposium "Neural Computation: From Perception to Cognitive Function", BCCN, Berlin, BBAW, Oktober 2012.

Workshop EU-Projektmanagement im 7. Forschungsrahmenprogramm, Charité Berlin, Dezember 2012.

Workshop "Compound Eyes: From biology to technology", Tübingen, März 2013).

Third EUCogIII Members Conference, Palma de Mallorca, April 2013.

IEEE/RSJ IROS'13 International Workshop on Vision-based Closed-Loop Control and Navigation of Micro Helicopters in GPS-denied Environments, November 2013, Tokyo, Japan.

2nd RED-UAS 2013 Workshop on Research, Education and Development of Unmanned Aerial Systems, November 2013, Compiegne, France.

Workshop des Humboldt-Princeton Centre for Reality Mining of Animal-Human Systems, Juli 2014.

13. Simulation of Adaptive Behavior 2014 Konferenz in Castellon, Spanien, Juli 2014.

Autonomous Learning Summer School 2014 (DFG), Leipzig, September 2014.

Workshop des Humboldt-Princeton Centre for Reality Mining of Animal-Human Systems, Dezember 2014.

Workshop "FETOpen für Antragstellerinnen und Antragsteller" der NKS des BMBF, Juni 2015.

Eigene Publikationen:
O. BERTHOLD, V. V. HAFNER, (2013), *Increased Spatial Dwell-Time for a Hovering Platform Using Optical Flow,* Technical Report.

O. BERTHOLD, V. V. HAFNER, (2013), *Unsupervised learning of camera exposure control using randomly connected neural networks,* 2nd RED-UAS 2013 Workshop on Research, Education and Development of Unmanned Aerial Systems, November 20-22, 2013, Compiegne, France.

O. BERTHOLD, V. V. HAFNER, (2013), *Neural sensorimotor primitives for vision-controlled flying robots,* IEEE/RSJ IROS'13 International Workshop on Vision- based Closed-Loop Control and Navigation of Micro -Helicopters in GPS-denied Environments, November 7, 2013, Tokyo, Japan.

C. BLUM, O. BERTHOLD, P. RHAN, AND V. V. HAFNER, (2014), *Intuitive control of small flying robots.* In Proceedings of the 2014 ACM/IEEE international conference on

Human-robot interaction (HRI '14), ACM, New York, NY, USA, 128-129. DOI=10.1145/2559636.2559826, http://doi.acm.org/10.1145/2559636.2559826

O. BERTHOLD AND V. V. HAFNER, (2014) *Unsupervised learning of sensory primitives from optical flow fields.* In A. P. del Pobil et al., editors, From Animals to Animats 13, volume 8575 of Lecture Notes in Computer Science, pages 188-197. Springer International Publishing, 2014.

O. BERTHOLD, V. V. HAFNER, (2015), *Closed-loop acquisition of behaviour on the Sphero robot,* Accepted for Publication in the Proceedings of the European Conference on Artifical Life (ECAL) 2015 Conference, York, UK. MIT Press.

DANIEL CAGARA

Schneller ans Ziel durch Fahrzeug zu Fahrzeug Kommunikation

Die Kernaufgabe meiner Forschung ist die kooperative Verbesserung von Routenwahlen in Straßennetzwerken im Kontext der Fahrzeug zu Fahrzeug Kommunikation. Hierzu war es zunächst erforderlich den Terminus „Verbesserung" näher zu spezifizieren. In meiner Forschung definiere ich Verbesserung Im Sinne einer Annäherung an das globale Optimum. Konkret versuche ich die Gesamtkosten (in etwa die Summe der Fahrzeit) für die Gesamtheit aller Verkehrsteilnehmer zu optimieren, auch wenn es hierbei dazu kommen kann, dass einige wenige Verkehrsteilnehmer zugunsten der Allgemeinheit einen Nachteil in Kauf nehmen müssen. Diese Art des Optimums wird im wissenschaftlichen Kontext als das „System Optimum" bezeichnet und ist im Vergleich zu anderen (egoistischen) Optimierungsverfahren ein NP-schweres (also nach aktuellem Stand der Forschung algorithmisch nicht effizient lösbares) Problem.

Eine mögliche Lösung für komplexe Probleme wie die Systemoptimierung sind genetische Algorithmen, die sich an evolutionären Prozessen aus der Natur anlehnen und so heuristisch gute Lösungen in einem Lösungsraum finden. Zunächst habe ich die Anwendbarkeit von genetischen Algorithmen auf die Optimierungsprobleme in Straßennetzwerken untersucht und im Rahmen einer Publikation auf der GECCO 2014 in Vancouver vorgestellt.

Dieser Arbeit diente nur als Verifikation der Anwendbarkeit und machte für den Produktivansatz unrealistische Annahmen wie in etwa dass globales Wissen über die Vergangenheit und die Zukunft vorliegt. Dies ist natürlich in einem realen Straßennetz nicht der Fall. Die schwierige Aufgabe war es nun diesen Algorithmus unter realen Bedingungen funktionsfähig zu machen.

Die Kernidee die ich verfolgt habe war die verteilte Implementierung dieses Algorithmus. In der Literatur wurde gezeigt, dass die Berechnung vieler kleiner genetischer Algorithmen gepaart mit einem regelmäßigen Informationsaustausch untereinander eine bessere Performance besitzt als die Implementierung einer

großen Instanz eines genetischen Algorithmus mit der gleichen Rechenleistung. So kam die Idee auf, die Rechenleistung aller einzelnen Navigationsgeräte zu bündeln. In diesem Kontext berechnet jedes teilnehmende Navigationsgerät eine Instanz des Optimierungsproblems basierend auf dem eigenen, durch Fahrzeug zu Fahrzeug Kommunikation gesammelten, Wissens über die aktuelle Verkehrslage. Ein weiterer Austausch mittels Fahrzeug zu Fahrzeug Kommunikation sorgt dafür, dass die so genannte Migration, also der Austausch von Informationen unter den einzelnen Instanzen der genetischen Algorithmen, erfolgen kann.

Die Evaluation eines solchen Algorithmus ist mit Bordmitteln nicht möglich. Typischerweise benutzen die genetischen Algorithmen einen mikroskopischen Verkehrssimulator, um die zukünftige Entwicklung des Verkehrs unter der Prämisse einer bestimmten Routenzuweisung an alle Fahrzeuge im Verkehrsnetz vorherzusehen. So können dann gute Routenwahlen von schlechten Routenwahlen separiert werden. Leider sind die aktuellen, mikroskopischen Simulatoren sehr langsam in der Ausführung (unter anderem da jedes einzelne Fahrzeug in jedem Detail seiner Bewegung simuliert werden muss). Unter diesen Umständen ist eine effiziente Evaluation des Algorithmus nicht möglich gewesen. Die Lösung habe ich mit einem eigens für diesen Anwendungszweck implementierten Verkehrssimulator gefunden. Die Herausforderung war so viele Aspekte wie möglich zu abstrahieren aber dennoch verlässliche Prognosen über die zukünftige Entwicklung des Verkehrs zu gewährleisten. Meine Implementierung des Simulators basiert auf dem Queuing Modell von Gawron. Erweitert werden musste das Modell um die Kreuzungsdynamiken, welche sich in realen Straßennetzen wiederfinden und, je nach Kreuzungslogik (also etwa einer rechts-vor-links Regelung, einer Vorfahrtsregelung oder einer Ampel) zu verschiedenen Verkehrsmustern führen. Meine Implementierung ist von der Rechenintensität her um den Faktor 100 schneller als herkömmliche Simulatoren und erlaubt es nun den eingangs beschriebenen, verteilten Algorithmus zur Optimierung der Routenwahl in komplexen Straßennetzen in akzeptabler Zeit zu evaluieren.

Erste Tests haben gezeigt, dass die durchschnittliche Fahrzeit durch die Anwendung dieses Algorithmus zu besseren Ergebnissen führt als herkömmliche, aktuell im Produktiveinsatz befindliche Systeme. Evaluiert wurde in einem realistischen Szenario der Stadt Bologna, dessen OD Matrix (also die Beschreibung von Fahrten zwischen verschiedenen Start- und Zielpunkten) anhand aus ca. 630 Detektordaten gewonnenen Daten konstruiert wurde und somit eine realistische Verkehrsnachfrage wiederspiegelt.

Forschungsaufenthalte, Konferenzen, Workshops:
GECCO 2014, Vancouver, Juli 2014.

IEEE Vehicular Networking Conference (VNC), Kyoto, Japan, Dezember 2015

Eigene Publikationen:
CAGARA, DANIEL, BJÖRN SCHEUERMANN, AND ANA LC BAZZAN. *A Methodology to Evaluate the Optimization Potential of Co-ordinated Vehicular Route Choices.* Inter-Vehicle Communication (FG-IVC 2013) (2013): 11.

CAGARA, DANIEL, ANA LC BAZZAN, AND BJÖRN SCHEUERMANN. *Getting you faster to work: a genetic algorithm approach to the traffic assignment problem.* Proceedings of the 2014 conference companion on Genetic and evolutionary computation companion. ACM, 2014.

BAZZAN, ANA LC, DANIEL CAGARA, AND BJORN SCHEUERMANN. *An evolutionary approach to traffic assignment.* Computational Intelligence in Vehicles and Transportation Systems (CIVTS), 2014 IEEE Symposium on. IEEE, 2014.

HARTMUT LACKNER

Testing of Variant-Rich Systems

Die Ansprüche von Kunden an neue Produkte sind gewachsen. Produkte sollen genau auf die einzelnen Kundenwünsche zugeschnitten sein, sodass der Kunde genau die Funktionen erhält, die er benötigt und keine überflüssigen Funktionen bezahlt. Hersteller reagieren auf diese gestiegenen Ansprüche mit immer mehr Varianten in denen sie ihre Produkte den Kunden anbieten. Die Variantenvielfalt hat in solchem Maß zugenommen, dass selbst in Massen gefertigte Produkte heute als Unikate gebaut werden können. Insbesondere die deutsche Automobilindustrie hat erkannt, das Individualisierbarkeit ein bedeutendes Verkaufsargument ist. Aktuelle Modelle werden in so vielen Varianten angeboten und verkauft, dass nur in seltenen Fällen eine Variante ein zweites Mal gekauft und produziert wird.

Der gegenwärtig gewünschte Variantenreichtum stellt jedoch neue Herausforderungen an die bisherigen Entwicklungs- und Testmethoden. Entwickelte Module werden wiederverwendet um verschiedene Varianten zu realisieren. Jedoch können in der Regel nicht alle miteinander gleichermaßen kombiniert werden. Prozesse die sich der systematischen Entwicklung variantenreicher Systeme widmen bezeichnen wir als Product Line Engineering (PLE). Im PLE können allgemeine und variable Teile des Systems gekennzeichnet werden. Zusätzliche Kombinationsvorschriften spezifizieren unter welchen Bedingungen Module miteinander kombinierbar sind. Wie viele Entwicklungsmethoden, lässt sich auch PLE modellbasiert unterstützen. In der Vergangenheit haben sich Feature-Modelle etabliert, mit der wir die Menge der gültigen Varianten in einer Produktlinie beschreiben können. Jedes Feature stellt hierbei ein Merkmal für den Kunden dar,

welches ihm einen zusätzlichen Nutzen bietet. Zusätzliche Mapping-Modelle stellen eine Verbindung zwischen den einzelnen Features und den einzelnen Modulen her.

Analog zur Systementwicklung, steht auch die Qualitätssicherung vor neuen Herausforderungen: Module einer Produktlinie müssen zwar einzeln getestet werden, aber dabei kann es nicht bleiben. Auch ihr Zusammenwirken in fertigen Produkten muss hinreichend getestet sein. Der Testaufwand, um jede Variante separat zu testen, ist jedoch zu hoch. Zwar ist der Testaufwand durch eine automatisierte Testdurchführung reduzierbar und auch automatisiertes modellbasiertes Testdesign verringert den Aufwand weiter. Aber dennoch ist es meistens nicht möglich alle Varianten zu testen, sondern nur einen Teil derer. Zur Stichprobenauswahl wurden bisher nur Verfahren vorgeschlagen, die auf Basis von Feature-Modellen eine systematische Auswahl treffen. Diese Verfahren vernachlässigen aber das Laufzeitverhalten der einzelnen Module untereinander.

Zielsetzung des Promotionsprojektes ist es ein Stichprobenverfahren zu entwickeln, welches auch das Laufzeitverhalten der einzelnen Module in das Testdesign und die Stichprobenauswahl einbezieht. Außerdem soll der empirische Beweis erbracht werden, dass das hier entwickelte Verfahren die Fehleraufdeckungswahrscheinlichkeit steigert. Dies wird anhand mehrerer Beispiele gezeigt. Zur Bewertung der Fehleraufdeckungswahrscheinlichkeit von Produktlinientests existieren bislang keine Methoden. Entsprechende Grundlagen hierfür müssen im Promotionsprojekt gelegt werden.

Aktueller Stand der Arbeit
Die fachlichen Problemstellungen sind weitestgehend gelöst und auf einschlägigen Konferenzen und Workshops veröffentlicht. Erfreulicherweise ist zu beobachten, dass die internationale Community die behandelten Themen aufgreift und in den Kontext ihrer eigenen Arbeiten setzt. Momentan befinde ich mich in der Phase des Zusammenfassens meiner Ergebnisse in der Dissertation.

Vernetzung mit anderen Projekten des Graduiertenkollegs.
In Zusammenarbeit mit Martin Schmidt, ebenfalls Doktorand im Graduiertenkolleg, sind zwei gemeinsame Publikationen entstanden. Zum einen identifizieren und analysieren wir Fehlerarten im PLE und zum anderen präsentieren wir einen Prozess zum Messen der Fehleraufdeckungswahrscheinlichkeit von Produktlinientests.

Ebenso produktiv ist die Zusammenarbeit mit Stephan Weißleder und Florian Wartenberg vom Fraunhofer Institut FOKUS bzw. der Thales Group. Zusammen haben wir die Grundlagen zur modellbasierten Testgenerierung aus Produktlinienmodellen gelegt. Basierend auf diesen Ergebnissen haben wir ein Verfahren zur Verbesserung der Testqualität bzgl. der Fehleraufdeckungswahrscheinlichkeit entwickelt. Die mit Martin Schmidt entwickelten Verfahren haben wir hierfür angewandt.

Forschungsaufenthalte, Konferenzen, Workshops:
International Conference on Software Testing, Verification and Validation (ICST), USA, April 2014.

Halmstad Summer School on Testing (HSST), Schweden, Juni 2014.

18th International Software Product Line Conference (SPLC), Florenz, Italien, September 2014.

Teilnahme amTesting Workshop, Bremen, August 2014.

Teilnahme am 35. GI Fachgruppentreffen Test, Analyse und Verifikation von Software (TAV), Ingolstadt, 2013.

Eigene Publikationen:
HARTMUT LACKNER, MARTIN THOMAS, FLORIAN WARTENBERG, STEPHAN WEIßLEDER: *Model-Based Test Design of Product Lines: Raising Test Design to the Product Line Level.* IEEE International Conference on Software Testing, Verification and Validation (ICST), 10 pages, Cleveland, U.S.A., 4/2014.

HARTMUT LACKNER, MARTIN SCHMIDT: *Assessment of Software Product Line Tests. Software Product Line Analysis Tools* (SPLat), 8 pages, Florence, Italy, 09/2014.

HARTMUT LACKNER, MARTIN SCHMIDT: *Potential Errors and Test Assessment in Software Product Line Engineering.* MBT 2015: p. 57-72.

STEPHAN WEIßLEDER, FLORIAN WARTENBERG, HARTMUT LACKNER*: Automated Test Design for Boundaries of Product Line Variants.* ICTSS 2015.

HARTMUT LACKNER: *Model-based Product Line Testing: Sampling Configurations for Optimal Fault Detection.* SDL Forum 2015.

HARTMUT LACKNER: *k-Shortest Paths with Limited Over*lap, 2015 (under review)

TOBIAS RAWALD

Efficient Analysis of High-Dimensional Datasets

Das Promotionsprojekt beschäftigt sich mit der Analyse hoch-dimensionaler Daten am Beispiel der sogenannten Recurrence Quantification Analysis (RQA), einer Methode aus der Zeitreihenanalyse. Hierbei werden im Speziellen das Problem der effizienten Berechnung langer Zeitreihen sowie der effizienten Exploration großer Ergebnismengen adressiert.

Die RQA basiert auf der Extraktion multi-dimensionaler Vektoren aus einer Zeitreihe, beispielsweise Messungen einer Klimareihe. Die extrahierten Vektoren repräsentieren die rekonstruierten, sich über die Zeit verändernden Zustände des Systems, bspw. des Klimas der Erde. Um zu bestimmen, ob ein System widerkehrendes Verhalten aufzeigt, werden die Vektoren auf paarweise Ähnlichkeit überprüft. Hierzu kommen Metriken, wie beispielsweise die

Euklidische Distanz, zum Einsatz. Die Ergebnisse der Vergleiche werden in einer binären Ähnlichkeitsmatrix (Recurrence Matrix) erfasst. Visualisierungen dieser Matrix werden als Recurrence Plot bezeichnet (siehe Beispiel in Abb. 1).

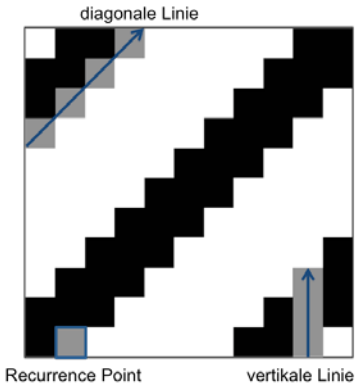


Abbildung 1: Recurrence Plot

Elemente der Matrix, die sich auf Paare von ähnlichen Vektoren beziehen, werden als Recurrence Points bezeichnet. Diese bilden kleinskalige Strukturen, insbesondere diagonale und vertikale Linien, die eine konkrete Semantik besitzen. Grundlage der RQA sind Histogramme dieser Linienstrukturen. Aus ihnen werden quantitative Maße abgeleitet. Hierzu zählt unter anderem der Determinismus, der den Anteil von Recurrence Points erfasst, die diagonale Linien formen.

Zur Durchführung der RQA steht eine Reihe von Software-Tools zur Verfügung. Hierzu zählen unter anderem die am Potsdamer Institut für Klimafolgenforschung entwickelte CRP Toolbox für Matlab sowie das Commandline Recurence Plots Tool. Die darunter liegenden Algorithmen basieren auf dem erschöpfenden Vergleich aller Paare von extrahierten Vektoren und liegen in der Komplexitätsklasse von O(n2). Die quadratische Zeitkomplexität erschwert die effiziente Analyse von sehr langen Zeitreihen, d.h. Zeitreihen mit mehr als einer Million Datenpunkten. Darüber hinaus liegt in vielen Fällen eine Limitierung bezüglich der Größe der verarbeitbaren Ähnlichkeitsmatrizen vor.

Das Ziel des Promotionsprojektes ist es zum einen, Ansätze zu entwickeln, die die mit RQA verbundenen Berechnungen effizienter gestalten. Dies soll es ermöglichen, sehr lange Zeitreihen zu analysieren. Darüber hinaus soll aufbauend ein Visual Analytics-Ansatz entwickelt werden, der die Menge der aus einer sehr langen Zeitreihe abgeleiteten Zeitreihen effizient zu analysieren. Zu diesem Zweck soll die RQA mit Multi-Skalen-Methoden, bspw. der Wavelet-Transformation, kombiniert werden.

Aktueller Stand

Bis zum jetzigen Zeitpunkt lag der Fokus der Arbeit auf der effizienten Berechnung sehr langer Zeitreihen. Im Folgenden werden die diesbezüglichen Fortschritte dokumentiert.

Matrix-Partitionierung

Um das Problem der Verarbeitung sehr langer Zeitreihen zu adressieren, wurde in einem ersten Schritt ein Ansatz zur Partitionierung der Recurrence Matrix entwickelt, der auf dem Konzept Divide & Recombine basiert. Hierbei wird die globale Recurrence Matrix in eine Menge von Teilmatrizen aufgeteilt (siehe Abb. 2). Für jede der Teilmatrizen werden lokale Histogramme der diagonalen und vertikalen Linienlängen bestimmt. Diese werden nach der Beendigung der Verarbeitung einer Teilmatrix in die entsprechenden globalen Histogramme überführt.

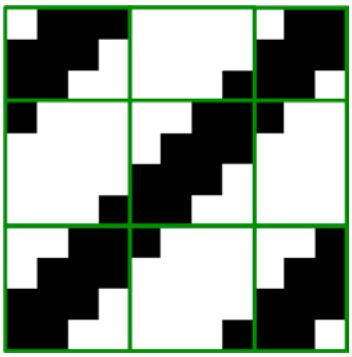


Abbildung 2: Matrix-Partitionierung

Aufgrund der Partionierung der globalen Matrix besteht die Möglichkeit, dass sich diagonale und vertikale Linien über eine Menge von Teilmatrizen erstrecken. Um zu gewährleisten, dass die RQA-Ergebnisse dennoch korrekt sind, führt der Ansatz zusätzliche globale Datenstrukturen ein. Sogenannte Carryover Buffer speichern die Länge der Linien an den horizontalen und vertikalen Grenzen der Teilmatrizen. Die zwischengespeicherten Ergebnisse dienen dabei als Input für die Detektierung von Linien in den angrenzenden Teilmatrizen.

Effiziente Verarbeitung

Bei der Entwicklung der Ansätze zur effizienten Durchführung der RQA wird grundsätzlich zwei Stoßrichtungen gefolgt:

1. *Parallele Verarbeitung*: Der Einsatz von paralleler Verarbeitung um die Rechenlast der RQA auf mehrere Berechnungseinheiten zu verteilen.
2. *Baumstrukturen*: Der Einsatz von Baumstrukturen, wie bspw. des k-d-Baums, um die Anzahl der durchgeführten Ähnlichkeitsvergleiche drastisch zu reduzieren.

Parallele Verarbeitung

Parallele Verarbeitung wird auf unterschiedlichen Ebenen eingesetzt. Einerseits bei der Prozessierung innerhalb einer Teilmatrix. Diese lässt sich grundsätzlich in drei Operatoren unterteilen:

1. ERZEUGE_MATRIX: Die Berechnung der paarweisen Ähnlichkeiten zwischen multi-dimensionalen Vektoren.

2. DETEKTIERE_DIAGONALE_LINIEN: Die Detektierung von diagonalen Linien innerhalb der Diagonalen der Teilmatrix.
3. DETEKTIERE_VERTIKALE_LINIEN: Die Detektierung von vertikalen Linien innerhalb der Spalten der Teilmatrix.

Jedem dieser Operatoren lassen sich atomare Ausführungseinheiten zuordnen. Hierbei ist die Bearbeitung einer Ausführungseinheit vollständig unabhängig von allen anderen Ausführungseinheiten desselben Operators. Es ergibt sich für jeden Operator ein bestimmter maximaler Grad paralleler Verarbeitung; basierend auf der Anzahl der extrahierten Vektoren n:

1. ERZEUGE_MATRIX: Ein Paar von multi-dimensionalen Vektoren (n2)
2. DETEKTIERE_DIAGONALE_LINIEN: Eine Diagonale der Teilmatrix (2n - 1)
3. DETEKTIERE_VERTIKALE_LINIEN: Eine Vertikale der Teilmatrix (n)

Hoch-parallele Berechnungsgeräte, wie bspw. GPUs, erlauben es, eine hohe Anzahl dieser Ausführungseinheiten nebenläufig auszuführen. Um diese Geräte anzusteuern, werden Implementierungen der parallelen Verarbeitung unter Verwendung des OpenCL-Frameworks bereitgestellt. Es erlaubt die Portierung von Quellcode zwischen Geräten unterschiedlicher Hersteller, was den Vergleich des Laufzeitverhaltens identischer Implementierungen ermöglicht. Hierbei wird deutlich, dass unterschiedliche Implementierungsstrategien auf unterschiedlicher Hardware eine unterschiedliche Performance liefern.

Die unterschiedlichen Implementierungsstrategien, lassen sich anhand der folgenden Dimensionen unterscheiden:

1. Repräsentation der Input-Daten,
2. Materialisierung der Ähnlichkeitsmatrix,
3. Repräsentation der Ähnlichkeitswerte und
4. Wiederverwendung von Zwischenergebnissen.

Anhand dieser Dimensionen wurden 5 Implementierungen mit unterschiedlichen Eigenschaften verglichen. Bezüglich der Evaluation ist ein konkretes Experiment dabei charakterisiert durch die folgenden Parameter:

1. Hardware-Plattform,
2. Implementierung,
3. Länge der extrahierten Vektoren und
4. Aktivierung/Nicht-Aktivierung der Default-OpenCL-Compiler-Optimierungen.

Die Experimente wurden auf einer CPU- und zwei GPU-Plattformen durchgeführt. Die Ergebnisse der Evaluation zeigen, dass unabhängig von der eingesetzten Hardware-Plattform, eine Implementierung mit folgenden Eigenschaften akzeptables Laufzeitverhalten aufweist:

1. Spaltenweise Repräsentation der Input-Daten,
2. Materialisierung der Ähnlichkeitsmatrix im Speicher des Berechnungsgerätes,
3. Byte-Repräsentation der Ähnlichkeitswerte und
4. Verzicht auf die Wiederverwendung von Zwischenergebnissen.

Zusätzlich zur parallelen Verarbeitung von Ausführungseinheiten innerhalb einer Teilmatrix, lassen sich mehrere Teilmatrizen ebenfalls nebenläufig verarbeiten. Voraussetzung hierfür ist, dass diese Teilmatrizen keine Ausführungseinheiten enthalten, zwischen denen Abhängigkeiten in der Prozessierung bestehen. Hierzu wurden Ausführungsebenen definiert, für die gilt, dass jede Teilmatrix der Ebene vollständig unabhängig von allen anderen Teilmatrizen derselben Ebene ist (siehe Abb. 3). Alle Teilmatrizen einer Ebene können hierbei nebenläufig verarbeitet werden. Die Ebenen selbst werden in aufsteigender Reihenfolge des Index verarbeitet.

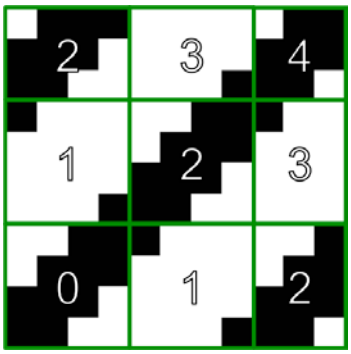


Abb. 3: Ausführungsebenen

Die Kombination der parallelen Verarbeitung innerhalb einer Teilmatrix unter Verwendung von hoch-parallelen Berechnungsgeräten sowie die nebenläufige Verarbeitung mehrerer Teilmatrizen ermöglicht die effiziente Verarbeitung sehr langer Zeitreihen. Am konkreten Beispiel der Potsdamer Reihe, einer Klimareihe mit mehr als einer Million Datenpunkten, konnte die Laufzeit der RQA von über 6 Stunden auf unter 100 Sekunden reduziert werden.

Baumstrukturen

Zusätzlich zum Einsatz der parallelen Verarbeitung bieten Baumstrukturen Optimierungspotential bei der Bestimmung der paarweisen Ähnlichkeiten der multi-dimensionalen Vektoren. Bei steigender Dimensionalität stellen die damit verbundenen Berechnungen einen zentralen Bestandteil des Gesamtaufwands dar. Das Ziel des Einsatzes von Baumstrukturen ist es dabei, die Anzahl der tatsächlich durchgeführten Ähnlichkeitsvergleiche drastisch zu reduzieren.

Zentrale Fragestellungen bezüglich des Einsatzes von Baumstrukturen sind:

1. Welche Baumstrukturen eignen sich für die RQA?
2. Existieren Bedingungen, unter denen Baumstrukturen bei der Bestimmung der

paarweisen Ähnlichkeiten ein besseres Laufzeitverhalten aufzeigen als parallele Implementierungen?
3. Besteht die Möglichkeit, den Einsatz der Baumstrukturen mit Elementen der parallelen Verarbeitung zu verknüpfen?

Zur Beantwortung dieser Fragen wurde in einem ersten Schritt ein Projekt im Rahmen einer studentischen Studienarbeit durchgeführt. Das Thema der Arbeit war der Vergleich von unterschiedlichen Implementierungen der zur RQA ähnlichen k-Nächsten Nachbar Suche. Der Fokus der Arbeit lag dabei auf dem Vergleich von parallelen Implementierungen zum Einsatz eines k-d-Baums. Die Ergebnisse der Arbeit zeigen, dass der k-d-Baum sensitiv gegenüber der Verteilung der Vektoren im Raum ist, im Gegensatz zu den parallelen Implementierungen. Darüber hinaus bietet er bei niedrigen Dimensionalitäten Einsparungspotential bezüglich der Laufzeit.

Ausgehend von diesen Ergebnissen wurden zwei RQA-Implementierungen auf Basis des k-d-Baums angefertigt, deren Laufzeitverhalten im Detail zu analysieren ist.

Ausblick
Im folgenden Schritt soll ein umfassendes Evaluierungsframework entwickelt werden, dass den Einfluss aller relevanten Parameter in Bezug auf das Laufzeitverhalten von RQA-Implementierungen erfasst. Zentrale Herausforderungen bilden dabei ein angemessenes Sampling des Untersuchungsraums sowie die Aufbereitung der Messergebnisse.

Die effizienten Implementierungen der RQA ermöglichen die Analyse sehr langer Zeitreihen. Dies dient als Grundlage, um die RQA mit Multi-Skalen-Methoden, im Speziellen der Diskreten Wavelet Transformation, zu verknüpfen. Ein weiteres Ziel des Promotionsprojektes ist es, einen Visual Analytics-Ansatz zu entwickeln, der es erlaubt, eine hohe Anzahl von RQA-Ergebnissen über verschiedene temporale Skalen effizient zu explorieren.

Forschungsaufenthalte, Konferenzen, Workshops:
Gemeinsamer Workshop der Informatik-Graduiertenkollegs Deutschlands, Schloss Dagstuhl, Juni 2015.

EDBT/ICDT 2015 Joint Conference (EDBT/ICDT) (März 2015) (Absage der Teilnahme aufgrund von Krankheit)

Eigene Publikationen:
RAWALD, T., SIPS, M., MARWAN, N., LESER, U. (2015): *Massively Parallel Analysis of Similarity Matrices on Heterogeneous Hardware.* - In: Fischer, P. M., Alonso, G., Arenas, M., Geerts, F. (Eds.), Proceedings of the Workshops of the EDBT/ICDT 2015 Joint Conference (EDBT/ICDT), (CEUR Workshop Proceedings ; 1330), p. 56-62.

RAWALD, T., SIPS, M., MARWAN, N., DRANSCH, D. (2014): *Fast Computation of Recurrences in Long Time Series.* - In: Marwan, N., Riley, M., Guiliani, A., Webber, C.

(Eds.), Translational Recurrences. From Mathematical Theory to Real-World Applications, (Springer Proceedings in Mathematics and Statistics ; 103), p. 17-29.

RAWALD, T., SIPS, M., MARWAN, N., DRANSCH, D. (2014): *Fast Recurrence Quantification Analysis on GPUs.* - In: NOLTA 2014 - International Symposium on Nonlinear Theory and its Applications, p. 325-329.

RAWALD, T., SIPS, M., MARWAN, N., DRANSCH, D. (2014): *Fast computation of recurrences in long time series,* (Geophysical Research Abstracts, Vol. 16, EGU2014-14824, 2014), General Assembly European Geosciences Union (Vienna 2014).

LARS GEORGE

Pattern Mining for Complex Event Processing

Complex Event Processing (CEP) Systeme finden Übereinstimmungen von Event Pattern in einem Ereignisstrom. Event Pattern beschreiben Zusammenhänge und Einschränkungen für passende Ereignisse und beschreiben so komplexe Situationen. Das Formulieren von Event Pattern ist nicht immer einfach, vor allem in Situation in der eine große Anzahl von Ereignissen kontinuierlich erzeugt wird. Selbst Domänenexperten können nicht immer die benötigten Zusammenhänge erkennen, die nötig sind um die gewünschte Situation durch ein Pattern zu beschreiben. Meine Arbeit beschäftigt sich mit der Aufgabe Event Pattern aus bestehenden Aufzeichnungen von Ereignissen zu erlernen.

In meiner Arbeit nutze ich Aufzeichnungen von Ereignissen in denen klar ist, dass die Situation die es durch ein Event Pattern zu beschreiben gilt vorgekommen ist. Im Forschungsbereich des maschinellen Lernens wird dies als überwachtes Lernen beschrieben. Traditionelle Techniken des maschinellen Lernens wie ein Neuronales Netz oder eine Support Vektor Maschine werden nicht genutzt, da die erlernten Modelle nicht direkt interpretierbar sind. Das Ergebnis meiner Algorithmen ist hingegen ein Event Pattern, welches im Klartext Reihenfolgen und Beschränkungen auf Ereignisse in einem Ereignisstrom darstellt.

Das Erlernen des Event Patterns unterteile ich in verschiedene Schritte, in denen jeweils ein Teil des Event Patterns erlernt wird. So werden die relevanten Ereignistypen, die Reihenfolge von Ereignissen, Beschränkungen der Eigenschaften der Ereignisse auf Gleichheit, Ungleichheit und Beziehungen zwischen Eigenschaften der Ereignisse in der Reihenfolge erlernt. Zusätzlich wird die maximal erlaubte Zeitspanne zwischen dem Eintreffen des ersten Ereignisses und des letzten Ereignisses des Event Patterns erlernt.

Erlernte Event Pattern werden in einem generischen Format gespeichert. Dieses kann im Anschluss in existierende CEP Sprachen transformiert werden. In meiner Arbeit habe ich dies bereits für die frei verfügbare open source Software Esper und die von Dr. Bruno Cadonna (ebenfalls in METRIK) entwickelte Sprache SES

beispielhaft implementiert. Erste Ergebnisse für Recall und Precision in Bezug auf die gefundenen Situationen von Interesse sind vielversprechend.

Ich plane eine ausführliche Evaluierung der entwickelten Algorithmen an synthetischen und realen Daten durchzuführen. Die synthetischen Daten möchte ich dafür nutzen, eine Aussage über die Möglichkeiten der zu findenden Pattern zu erhalten. So kann beispielsweise der Ansatz in Bezug auf die Eigenschaft des Patterns wie die Länge der vorkommenden Sequenzen oder Anzahl von Beziehungen zwischen Ereigniseigenschaften bewertet werden. Desweiten möchte ich weitere, komplexere Beschränkungsbeschreibungen für das Event Pattern erlernen können und somit meinen Ansatz erweitern.

Forschungsaufenthalte, Konferenzen, Workshops:
Gemeinsamer Workshop der Informatik-Graduiertenkollegs Deutschlands, Schloss Dagstuhl, Juni 2014.

Eigene Publikationen:
L. GEORGE: *Event Pattern Mining for Smart Environments,* In Proceedings of the 17th International SDL Forum (SDL 2015), Berlin, Germany, October 12-14, 2015, Lecture Notes in Computer Science (LNCS), Vol. 9369, Springer (2015)

L. GEORGE: *Automated Pattern Learning for Event Pattern Matching*, Proceedings of 9th Joint Workshop of the German Research Training Groups in Computer Science, Dagstuhl, Germany, June 16th-18th (2015)

FABIAN FIER

Effiziente Multi-Domain Ähnlichkeitssuche auf großen Daten

Ich beschäftige mich in meiner Forschung mit der Ähnlichkeitssuche auf großen Daten, wie sie z. B. beim Crawlen des Webs anfallen. In der Literatur wird dieses Thema bisher hauptsächlich im Zusammenhang mit einer einzigen Domäne, z. B. Text, beleuchtet. Anwendungsgebiete einer solchen Ähnlichkeitssuche sind z. B. die Deduplizierung von Textdokumenten, Plagiatserkennung oder die Erkennung von Einbruchsversuchen in Netzwerke. In der realen Welt werden Objekte aber häufig durch komplexere Daten abgebildet: Daten haben üblicherweise nicht nur eine relevante Domäne, sondern mehrere, z. B. Text, Ort und Zeit. Beispiele für solche komplexe Daten sind benutzergenerierte Daten von mobilen Geräten oder durch strukturierte Metadaten (schema.org) versehene Webseiten. Eine Ähnlichkeitssuche über mehrere Domänen liefert die passendsten Treffer, welche textuell ähnlich, vom Ort her nah und von der Zeit her nah sind. Anwendungsbereiche von Ähnlichkeitssuchen über solche komplexen Daten sind z. B. die Koordination von Hilfe in Krisengebieten, in denen benutzergenerierte Daten mit Hilfe von mobilen Endgeräten entstehen. Auch die Websuche, welche bisher als reine Textsuche ausgeführt ist, kann durch eine Ähnlichkeitssuche weiterentwickelt werden.

Ähnlichkeitssuchen über mehrere Domänen auf großen Daten sind bislang nicht systematisch erforscht.

Mein Ansatz ist, zunächst Algorithmen für Ähnlichkeitsjoins so weiterzuentwickeln, dass sie mehrere Domänen unterstützen. Ähnlichkeitsjoins berechnen alle Paare von ähnlichen Elementen über eine Eingabemenge von Elementen. Diese Joinalgorithmen dienen mir in einem zukünftigen Schritt als Grundlage für neuartige Algorithmen für Ähnlichkeitssuche.

Im Bereich textueller Ähnlichkeitsjoins existiert bereits eine Reihe von Algorithmen, deren Eigenschaften bislang nicht experimentell verglichen worden sind. Ich habe verschiedene relevante Algorithmen zusammengetragen und auf MapReduce implementiert. Aktuell vergleiche ich deren Eigenschaften, insbesondere hinsichtlich Laufzeit und Skalierbarkeit in Abhängigkeit von spezifischen Eigenschaften der Eingangsdaten, z. B. Wortverteilung und Dokumentlängen. Der Vergleich ist schon deshalb nicht trivial, da eine Vielzahl von Vergleichsdimensionen bestehen, welche für die Laufzeiteigenschaften und Skalierbarkeit der zu vergleichenden Algorithmen maßgeblich sind, z. B. Self-Join vs. RxS-Join, verschiedene textuelle Ähnlichkeitsmaße, Mengen vs. Multimengen/ Dokumente, Vorhandensein von Stop-Worten (besonders häufig auftretende Worte), Parametrisierung der Algorithmen und Parametrisierung der Ausführungsumgebung. Der in Bezug auf bestimmte Dateneigenschaften optimale Algorithmus soll als Grundlage für die Entwicklung neuer Algorithmen für Ähnlichkeitsjoins dienen, welche zusätzlich die Domäne Ort berücksichtigen können. Dies ist eine skalierbare Weiterentwicklung der Arbeit von „Spatio-textual similarity joins“ (Bouros et al.).

Die entstehenden Algorithmen verwende ich in einem zukünftigen Schritt, um Ähnlichkeitssuche zu ermöglichen. Hierbei plane ich, die Zwischenergebnisse der Joins zu persistieren, um Abfragen darüber ausführen zu können. Die Zwischenergebnisse können hybride Indexe sein, welche z. B. aus einem invertierten Index und einem Grid bestehen oder auch kombinierte Hash-Werte bzw. Signaturen. Eine Abfrage über einen solchen Index liefert eine Übermenge der Ergebnismenge. Durch einen Verifikationsschritt wird das Ergebnis daraus berechnet.

Meine Vision ist die Entwicklung eines möglichst generischen Frameworks von algorithmischen Ansätzen, um damit eine skalierbare performante Ähnlichkeitssuche auf komplexen großen Daten zu ermöglichen.

Forschungsaufenthalte, Konferenzen, Workshops:
Fachtagung Datenbanksysteme für Business, Technologie und Web (BTW), Hamburg, 2015.

Conference on Very Large Databases (VLDB), Riva del Garda, 2013.

Gemeinsamer Workshop der Informatik-Graduiertenkollegs Deutschlands, Schloss Dagstuhl, Juni 2015

Eigene Publikationen:

F. FIER AND P. BOUROS, *Textual Similarity Joins on MapReduce: An Experimental Evaluation. 2015.* (under preparation)

Spatio-Textual Similarity Joins on MapReduce. In Bearbeitung, Einreichung voraussichtlich Anfang 2016.

HEINRICH MELLMANN

Aktive Räumliche Situationsmodellierung für Autonome Mobile Roboter

Mit dem Fortschritt der Forschung werden die Roboter in immer mehr Bereichen eingesetzt. In der nahen Zukunft werden Roboter als Mitglieder von Smart-Cities oder Helfer in Katastrophenszenarien einen regulären Platz haben. Man stelle sich eine Gruppe heterogener autonomer Roboter vor, etwa fliegende Drohnen, fahrende Einheiten, etc., die ein Katastrophengebiet selbständig explorieren und ein schnelles und genaues Bild der Lage liefern. Um diese Aufgabe lösen zu können muss jeder Roboter in der Lage sein eine adäquate Räumliche Repräsentation der Situation zu erstellen und zu pflegen, die es erlaubt Resultate geplanten Aktionen vorherzusagen, d.h., zu planen, und die Resultate bereits ausgeführter Aktionen zu erklären, d.h. Ergebnisse früherer Entscheidungen bewerten. Um es zu erreichen muss der Roboter aktiv Information über seine Umgebung sammeln und integrieren sowie nicht direkt beobachtbare Informationen ableiten.

Unvollständige und verrauschte Sensorische Messungen führen zu Unsicherheiten in der Wahrnehmung des Roboters. Insbesondere falsche Wahrnehmungen (false perceptions) und Mehrdeutigkeit machen die Aufgabe zu einer Herausforderung. Verrauschte oder falsche Wahrnehmungen können zu Inkonsistenzen in der Wahrnehmung führen, während redundante Informationen von verschiedenen Sensoren dazu genutzt werden können um die Unsicherheit zu reduzieren. Alle diese Aspekte wurden zwar in verwandten Gebieten, wie etwa Signalverarbeitung, ausführlich studiert, jedoch stellt ein autonomes System in einer unbeschränkten dynamischen Umgebung eine gänzlich andere Ebene der Komplexität.

Viele aktuelle Ansätze beschränken sich oft auf statische, stark strukturierte Szenarien, meistens mit einer fahrenden Plattform, etwa ein Haushaltsroboter. Dabei werden viele Annahmen über die Umwelt möglich, z.B. waagerechte Lage der Kamera, wenige zeitliche Einschränkungen. So entstehende Lösungen versagen in einer komplexen dynamischen Umgebung, wo unter Zeitdruck und hoher Unsicherheit Entscheidungen getroffen werden müssen.

Insbesondere fokussieren sich Bemühungen oft auf einzelne Aspekte, die dann modular nach black-box prinzip zusammengesteckt werden. Wodurch keine geschlossenen Lösungen entstehen und viel Potential verloren geht.

Dieses Projekt hat zum Ziel ein dynamisches, geschlossenes Modell für räumliche und Zeitliche Wahrnehmung für autonome mobile Roboter zu entwickeln. Das Modell soll nach dem Vorbild des Menschen soll das Model die vier zentralen Aspekte Aufweisen: sensorische Wahrnehmungen integrieren um Abstraktere Information zu extrahieren; Interpretieren der Sensordaten basierend auf der Information aus höheren Abstraktionsstufen; relevante Daten auswählen (Aufmerksamkeit); direkte Inferenz von Entscheidungen ermöglichen. Dabei soll die Wahrnehmung in einer geschlossenen Schleife zwischen Sensoren und Aktuatoren ausgeführt werden, wo die Auswirkungen von Aktionen explizit in das Modell einfließen.

Als Experimentalumgebung wurde der humanoide Roboterfußball RoboCup gewählt. In diesem Szenario spielen Teams humanoider Roboter autonom gegeneinander Fußball. Jeder der Roboter ist mit einer Anzahl unterschiedlicher heterogener Sensoren ausgestattet, wie etwa Beschleunigungssensoren, Videokameras, Kraftsensoren an den Füßen etc. Roboter spielen in Teams, aber jeder muss in der Lage sein sich zu orientieren. Damit bietet RoboCup eine hervorragende Umgebung um den vorgeschlagenen Ansatz zu testen. Insbesondere kann die Leistung der Verfahren gegen alternative Ansätze anderer Gruppen unter standardisierten Bedingungen getestet werden.

Arbeit mit realen Robotern erfordert einen erheblichen infrastrukturellen Aufwand (Simulation, Logging, Programmarchitektur, etc.). Großer Teil notwendiger Infrastruktur wurde bereits entwickelt und getestet. Zusätzlich wurde über einen Zeitraum von etwa zwei Jahren eine große Basis An Sensordaten synchronisiert mit Videos aus realen Spielen gesammelt, was für Empirische Analyse der Wahrnehmung verwendet werden soll.

Darüber hinaus wurden verschiedene Teilaspekte der Wahrnehmung implementiert und untersucht, wie etwa Modellierung eigener Position und anderer Aspekte der Umgebung. Dabei wurde eine Reihe verschiedener Methoden eingesetzt und auf Anwendbarkeit untersucht, darunter sind Coonstraint-Basierte Verfahren, Probabilistische Verfahren wie Partikel Filter und Multi-Hypothesen Kalman Filter. Inferenz von Entscheidungen basierend auf der probabilistischer Wahrnehmung, sowie unterschiedliche dynamische Aktionen (stabiles Laufen, Ball Schießen, Dribbeln) wurden ebenfalls entwickelt und bei Wettbewerben eingesetzt.

Die bereits implementierten Teilaspekte wurden erfolgreich bei internationalen Wettbewerben eingesetzt und bei verschiedenen Konferenzen veröffentlicht (siehe Publikationsliste). Besonders haben sich die Verfahren basierend auf Patrikel-

simulation und Repräsentation der Daten als Graphen bewährt. Basierend auf dem gesammelten Wissen, soll das Gesamtmodell Struktur eines Hierarchischen Graphen aufweisen mit Partikelsimulation als Integrations- und Inferenzmethode.

In der letzten Phase des Projekts sollen die einwickelten Teillösungen in einem Gesamtmodell zusammengefasst und einer empirischen Auswertung unterzogen werden. Entwicklung eines geeigneten Testverfahrens um die Leistung des Models bewerten zu können. Da es sich um ein geschlossenes System (closed-loop) handelt, müssen die Tests auf einer realen Plattform durchgeführt werden.

Forschungsaufenthalte, Konferenzen, Workshops:
RoboCup 2015 in Hefei, China

RoboCup 2014 in João Pessoa, Brasilien.

RoboCup 2013 in Eindhoven.

Gemeinsamer Workshop der Informatik-Graduiertenkollegs Deutschlands, Schloss Dagstuhl, Juni 2015.

14th IEEE-RAS International Conference on Humanoid Robots (Humanoids), Madrid, Spain, November 2014.

22nd International Workshop on Concurrency, Specification and Programming (CS&P 2013).

Eigene Publikationen:
SCHEUNEMANN, M. M. & MELLMANN, H., *Multi-Hypothesis Goal Modeling for a Humanoid Soccer Robot*, in 'Proceedings of the 9th Workshop on Humanoid Soccer Robots, 14th IEEE-RAS International Conference on Humanoid Robots (Humanoids), (2014).

MELLMANN, H.; SCHEUNEMANN, M. & STADIE, O., *Adaptive Grasping for a Small Humanoid Robot Utilizing Force- and Electric Current Sensors*, in Marcin S. Szczuka; Ludwik Czaja & Magdalena Kacprzak, ed., 'Proceedings of the 22nd International Workshop on Concurrency, Specification and Programming (CS&P)', CEUR-WS.org, Warsaw, Poland, pp. 283-293, (2013).

KADEN, S.; MELLMANN, H.; SCHEUNEMANN, M. & BURKHARD, H.-D., *Voronoi Based Strategic Positioning for Robot Soccer*, in Marcin S. Szczuka; Ludwik Czaja & Magdalena Kacprzak, ed., 'Proceedings of the 22nd International Workshop on Concurrency, Specification and Programming (CS&P)', CEUR-WS.org, Warsaw, Poland, pp. 271-282, (2013).

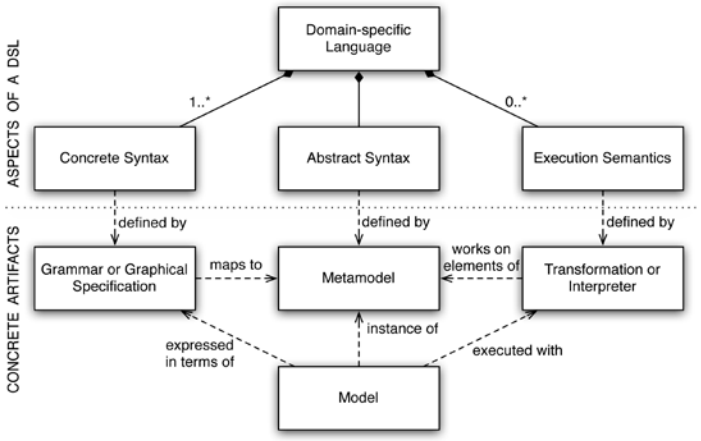
MARTIN SCHMIDT

Coupled Evolution in Domain-Specific Language Development

Inhalt des Promotionsprojektes
Domänenspezifische Sprachen (DSLs) sind Computersprachen, die auf eine Anwendungsdomäne spezialisiert sind. Dabei ermöglichen sie entgegen

Hochsprachen (GPLs), auch Domänenexperten, die nicht zwangsläufig Programmierer sind, Software zu entwickeln. Grund hierfür ist die Verwendung von domänenspezifischen Sprachkonzepten. Die Benutzergruppen von DSLs sind häufig kleiner als die der GPLs. Aufgrund dessen muss hier eine Abwägung zwischen Kosten und Nutzen für die Entwicklung einer DSL erfolgen. Zudem sind DSLs aufgrund ihrer engen Bindung zu den Domänenexperten häufig Änderungen bzw. Erweiterungen ausgesetzt.

Eine DSL besteht aus drei Aspekten, die durch unterschiedliche Formalismen beschrieben werden. (1) Die abstrakte Syntax beschreibt die Domänenkonzepte und wird im Bereich der modellgetriebenen Entwicklung durch ein Metamodell beschrieben. (2) Die konkrete Syntaxbeschreibung definiert mit welchen Sprachkonstrukten Modellinstanzen beschrieben werden. (3) Die Ausführungssemantik definiert die Ausführbarkeit der Modellinstanzen. Dies kann durch die Implementierung von Interpretern oder Generatoren erfolgen. Zwischen diesen Aspekten bestehen Abhängigkeiten, da die konkrete Syntaxbeschreibung und Ausführungssemantiken Elemente des Metamodells referenzieren (siehe Grafik).



Für den DSL-Entwickler bedarf es deshalb verschiedener Werkzeuge, die es erlauben Änderungswünsche kostengünstig und fehlerfrei entlang dieser Abhängigkeiten durchzuführen, um eine iterative Entwicklung einer DSL in Zusammenarbeit mit den Domänenexperten zu ermöglichen. Änderungen, die sich auf Bestandteile eines Systems auswirken, haben Änderungen an anderen Bestandteilen zufolge. Diese Änderungen werden als Koadaptionen bezeichnet. Ziel dieser Arbeit ist es, Werkzeuge in existierende Editoren für die DSL-Entwicklung im modellgetriebenen Kontext zu integrieren, welche Koadaptation (semi-)automatisiert ermöglichen.

Im ersten Teil meiner Arbeit beschäftige ich mich mit der Analyse von Kombinationen an Sprachen, die in der modellbasierten DSL-Entwicklung verwendet werden. Hierbei wird eine Kategorisierung von Abhängigkeiten

zwischen den Beschreibungssprachen vorgenommen. Darüber hinaus wird ein Vergleich von Änderungen in der DSL-Entwicklung zu Änderungen in der objektorientierten Programmierung (OOP) durchgeführt. Konkret bedeutet dies, ob Konzepte wie Refaktorisierungen, Konstruktionen oder Destruktionen auch in der modellbasierten DSL-Entwicklung angewendet werden können.

Für die Durchführung von Koadaptionen wird zunächst ein Koadaptionskatalog aufgestellt. Entsprechend dieses Kataloges werden geeignete Transformationstechniken ausgewählt, mithilfe derer Koadaptionen durchgeführt werden können. Hierfür erfolgt zudem die Konzeption und Entwicklung von Algorithmen, die es erlauben Abhängigkeiten zwischen den Sprachbestandteilen zu identifizieren. Die identifizierten Abhängigkeiten werden als abstrahierte Sicht für die jeweilige Koadaption genutzt, um die Änderungen mittels Transformationen durchzuführen.

Für die Integration dieser Änderungen soll im Rahmen meiner Arbeit eine eigene DSL entwickelt werden, welche es ermöglicht Refaktorisierungen zu beschreiben. Auf Basis dieser Beschreibungen sollen anschließend Erweiterungen für bestehende Editoren generiert werden. Abschließend sollen diese Erweiterungen auf eine Ansammlung von DSLs angewendet werden und deren Effizienz evaluiert werden.

Aktueller Stand der Arbeit

Die Recherchearbeit und Problemanalyse wurde abgeschlossen. Es wurden Untersuchungen auf bestehenden DSL-Projekten innerhalb des Graduiertenkollegs vorgenommen. Hierbei wurden die einzelnen Schritte in der Entwicklung einer DSL untersucht und die Änderungen zwischen zwei Schritten aufgezeichnet und ausgewertet. Hieraus haben sich Ähnlichkeiten zu Koadaptionen im Bereich der OOP ergeben. Basierend auf dieser Erkenntnis wurde vorläufig ein Adaptionskatalog aus dem Bereich der OOP als Ausgangslage für weitere Arbeiten gewählt. Zudem wurde dieser Katalog um Ergebnisse aus der Arbeit von Herrn Daniel Sadilek erweitert.

Für die Übertragung von Änderungen werden asymmetrische bidirektionale Modeltransformationen ausgewählt. Diese ermöglichen die Übertragung von Änderungen auf heterogenen Graphstrukturen. Dieser Ansatz wurde im Rahmen meiner Arbeit adaptiert und bereits als Beitrag bei einer Konferenz veröffentlicht. Aktuell beschäftige ich mich mit der Formalisierung und Implementierung zur Identifikation von Abhängigkeiten. Augenmerk hierbei liegt auf eine performante Lösung, die es DSL-Entwicklern ermöglicht, zur Entwicklungszeit direkt entlang der Abhängigkeiten zwischen den DSL-Bestandteilen zu navigieren.

Mit Herrn Arif Wider fand vor dem Eintritt in das Graduiertenkolleg eine Zusammenarbeit statt im Bereich der Entwicklung einer domänenspezifischen Werkzeugkette für den Einsatz in der Nanophysik. Diese Zusammenarbeit wurde

als Beitrag bei einer Konferenz eingereicht und diente als Motivation für das eigene Promotionsprojekt.

Als Resultat der Zusammenarbeit mit Herrn Hartmut Lackner entstanden zwei Veröffentlichungen im Bereich der Mutationsanalyse von Testmodellen für Softwareproduktlinien.

Forschungsaufenthalte, Konferenzen, Workshops:
Forschungsaufenthalt am University Residential Center of Bertinoro, Italien. Summer School on Formal Methods for the Design of Computer, Communication and Software Systems: Model-Driven Engineering.

MBT 2015, 10th Workshop on Model-Based Testing, London, April 2015.

Gemeinsamer Workshop der Informatik-Graduiertenkollegs Deutschlands, Schloss Dagstuhl, Juni 2015.

Eigene Publikationen:
HARTMUT LACKNER, MARTIN SCHMIDT*: Potential Errors and Test Assessment in Software Product Line Engineering* (15 Seiten). ETAPS: 10th Model-Based Testing Workshop. 2015.

HARTMUT LACKNER, MARTIN SCHMIDT: *Assessment of Software Product Line Tests* (8 Seiten). SPLat - Software Product Line Analysis Tools. 2014.

MARTIN SCHMIDT, ARIF WIDER, MARKUS SCHEIDGEN, JOACHIM FISCHER, SEBASTIAN VON KLINSKI: *Refactorings in Language Development with Asymmetric Bidirectional Model Transformations* (17 Seiten). SDL Forum: Model-Driven Dependability Engineering. 2013.

JOHANNES SCHREYER

Modellierung urbaner Vegetation mit fernerkundlichen Methoden auf verschiedenen räumlichen Skalen

Die globale Verstädterung bewirkt den Bedeutungsgewinn von Stadtvegetation und den damit assoziierten urbanen Ökosystemdienstleistungen, u.a. CO2 Sequestrierung von Bäumen. Neben flächenhaften Informationen über Zusammensetzung und Verbreitung von städtischen Grünflächen sind dreidimensionale Informationen für die Quantifizierung der Ökosystemdienstleistungen von Bedeutung. Techniken zur dreidimensionalen Erfassung von Gebäuden wurden in den vergangen Jahren kontinuierlich weiterentwickelt und haben Einzug in digitale Informationssysteme, wie Google Maps erhalten. Ansätze zur dreidimensionalen Parametrisierung urbaner Vegetationsbestände aus Bäumen, Gehölzen und Sträuchern im öffentlichen und privaten Raum sind dagegen gering entwickelt. Fernerkundungs-gestützte Erfassungen bieten dabei nicht nur eine kostengünstige, sondern auch flächendeckende Alternative zu terrestrischen Verfahren. Dieses Promotionsprojekt widmet sich deshalb der Evaluierung neuer fernerkundlicher

Techniken und Produkte für die Erfassung dreidimensionaler Parameter wie Vegetationshöhe, Volumen, Kronenbreite einzelner Vegetationsbestandteile und zusammenhängender Vegetationsbestände.

Ansatz 1:
Im ersten Teil der Arbeit wurde ein digitales Oberflächenmodell der TanDEM-X Mission genutzt, um ein normalisiertes Vegetationsmodell für typische Stadtvegetation (u.a. Straßenbäume) und waldähnliche Stadtvegetation (z.B. in Parks oder auf Friedhöfen) für eine Testfläche in Berlin zu erstellen. In Kollaboration mit dem Deutschen Luft- und Raumfahrtzentrum in Oberpfaffenhofen wurde ein mehrstufiges Verfahren entwickelt, dass bestehende mit neu entwickelten Ansätzen der mathematischen Morphologie und Objekt-basierten Bildbearbeitung verwendet. Mit einem eigenen innovativen Ansatz, der einen progressiven morphologischen Filter und ein disaggregiertes Oberflächenmodell als Grundlage nutzt, konnte für beide Stadtvegetationstypen die besten Ergebnisse erzielt werden. Es konnte nachgewiesen werden, dass mit dem entwickelten Ansatz eine Höhenabschätzung durch ein normalisiertes Kronenmodell möglich ist. Die Ergebnisse können als Grundlage für großräumige Studien genutzt werden, wie z.B. einer stadtübergreifenden Kohlenstoffabschätzung.

Ansatz 2:
Parallel zu diesem großräumigen Ansatz mit TanDEM-X Daten werden in Kollaboration mit Oswald Berthold, Doktorand des GRK METRIK aus der Arbeitsgruppe für kognitive Robotik des Institut für Informatik, HU Berlin, mehrere unbemannte Flugsysteme für die Erstellung dreidimensionaler Punktwolken kleinräumiger Stadtvegetation getestet. Dabei wurde bisher an der geeigneten Konfiguration eines Mikrokopters gearbeitet, der auch in beengten Räumen, wie unterhalb des Kronendachs, verortete und störungslose RGB Bilder von Vegetation erstellen kann. Aus den Aufnahmen kann über einen photogrammetrischen Workflow die räumliche Struktur der Vegetationsbestandteile in einer Punktwolke rekonstruiert werden. Ähnlich wie bei der Anwendung des TanDEM-X Oberflächenmodells kann aus der Punktwolke ein normalisiertes Vegetationsmodell erstellt werden, um daraus dreidimensionale Parameter, wie Höhe und Volumen zu gewinnen. Daneben wird in den kommenden Monaten ein bestehendes System ("Agricopter", Ferry Bachmann, Arbeitsgruppe für kognitive Robotik, HU Berlin) für die dreidimensionale Erfassung von Stadtvegetation und Differenzierung zu angrenzenden Gebäuden aus größerer Höhe getestet.

Forschungsaufenthalte, Konferenzen, Workshops:
Forschungsaufenthalt, Deutsches Luft- und Raumfahrtzentrum (DLR) Oberpfaffenhofen, März 2015.

JURSE Konferenz, Lausanne, Schweiz, März 2015.

Workshop EFRE Projekt „Stadtklima“.

ICCSA Konferenz, Guimaraes, Portugal, Juli 2014.

3rd workshop of the working group evaluation of remote sensing data, TU Berlin, Oktober 2014.

SURE World Conference, HU Berlin, Juli 2013.

Eigene Publikationen:

SCHREYER, J./ GEIß, C. / LAKES, T (2015). *TanDEM-X for large-area modeling of 3D urban vegetation – Evidence from Berlin,* Germany. Submitted to IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing (IEEE J-STARS).

SCHREYER, J./ LAKES, T (2015). *TanDEM-X & UAV data for modeling 3D vegetation information in urban areas* (*Proceeding & Poster*). Published in Proc. IEEE-CPS Joint Urban Remote Sensing Event (JURSE), Lausanne, Switzerland, March 2015.

SCHREYER, J./ TIGGES, J./ LAKES, T./ CHURKINA, G. (2014). *Using Airborne LiDAR and QuickBird Data for Modelling Urban Tree Carbon Storage and Its Distribution - A Case Study of Berlin.* Remote Sensing, 6, 10636-10655.

SCHREYER, J./ LAKES, T. (2014). *New remote-sensing based approaches for modeling 3D vegetation information for ESS analyses in urban areas* (*Proceeding*). Submitted and accepted to Proc. IEEE-CPS International conference on computional science and its applications, Guimareas, Portugal, July 2014.

SCHREYER, J. (2014). *Remote-sensing based approaches for modeling 3D vegetation information in urban areas* (Poster). 3rd workshop of the working group evaluation of remote sensing data by the DGPF, Berlin, Germany, October 2014.

Forschungsprojekte und bisher erzielte Forschungsergebnisse:

BRUNO CADONNA

Daten- und Ereignisstromverarbeitung

Ziel meiner Forschung ist der Entwurf, die Entwicklung und die Evaluierung neuer Methoden für das Abfragen von Daten- und Ereignisströmen. Meine Forschung kann in drei Bereiche gegliedert werden: Das Auffinden von Instanzen von Ereignismustern in Ereignisströmen, das Erlernen der Ereignismuster aus historischen Ereignissequenzen und die verteilte und parallele Datenstrom-verarbeitung.

Das Auffinden von Instanzen von Ereignismustern in Ereignisströmen (engl. Event Pattern Matching - EPM) ist eine Abfragetechnik, bei der Instanzen (Matches) eines Musters (Pattern) in einem Strom von Ereignissen (Events) gefunden werden. Ereignisse sind Datenelemente mit einem Zeitstempel, der den Zeitpunkt des Auftretens des Ereignisses beschreibt. Beispiele von Ereignissen sind

Verabreichungen von Medikamenten, Aktienverkäufe oder die Messungen von Sensoren. Ein Muster beschreibt ein zusammengesetztes Ereignis und beinhaltet Bedingungen hinsichtlich der zeitlichen Ordnung, der Attributwerte sowie der Anzahl und der zeitlichen Ausdehnung jener Ereignisse, die gemeinsam das zusammengesetzte Ereignis darstellen. Eine Instanz des Musters ist ein gesuchtes zusammengesetztes Ereignis im Ereignisstrom und besteht aus jenen Ereignissen, die die Bedingungen im Muster erfüllen. Abbildung 1 stellt EPM schematisch dar.

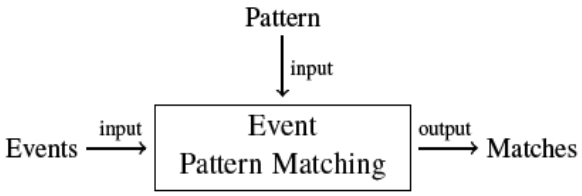


Abbildung 1: Event Pattern Matching

EPM kann in unterschiedlichen Domänen angewendet werden, wie z.B. RFID-basierter Standortverfolgung, RSS-Feeds Überwachung, medizinische Datenverarbeitung, Sensornetzwerke und Arbeitsablaufüberwachung. Beispielanwendungen für EPM im Kontext von SmartCity sind die Erkennung vorbeifahrender PKWs und LKWs aufgrund von Sensordaten und die Erkennung von interessanten Situationen (z.B. Ausbruch eines Feuers) in intelligenten Gebäuden. Die Erkennung vorbeifahrender PKWs und LKWs kann verwendet werden, um kurzfristig den Verkehr zu regeln und langfristig die Verkehrsplanung einer Stadt zu verbessern. Dabei beschreiben die Muster die typischen Sensormessungen (Ereignisse) während der zusammengesetzten Ereignisse "vorbeifahrender PKW" und "vorbeifahrender LKW". Durch die Erkennung von interessanten Situationen in Gebäuden, die mit Sensorennetzwerken ausgerüstet sind, können geeignete Maßnahmen zeitnah getroffen werden. Zum Beispiel kann der Anstieg der Temperatur mit gleichzeitiger Rauchentwicklung im gleichen Bereich des Gebäudes auf den Ausbruch eines Feuers hindeuten. Wird diese Situation erkannt, wird das Gebäude evakuiert und die Feuerwehr verständigt. Die Herausforderungen bei EPM sind, das Rauschen in Ereignisströmen zu vernachlässigen, die Latenz und den Speicherbedarf zu minimieren, sowie den Durchsatz zu maximieren.

Zusätzlich zum Muster kann durch eine Auswahlstrategie (engl. selection strategy) die Ergebnismenge der gefundenen Instanzen eines Musters in einem Ereignisstroms an anwendungsspezifische Anforderungen angepasst werden. Dadurch kann das speicherintensive, Durchsatz vermindernde und Latenz steigernde Auffinden aller möglichen Instanzen eines Musters in einem Ereignisstrom vermieden werden. Die in der Literatur vorgeschlagenen Auswahlstrategien begrenzen die Instanzen in der Ergebnismenge anhand der Positionen der

Ereignisse einer Instanz im Ereignisstrom zueinander. Zum Beispiel bestimmt die Auswahlstrategie „contiguous“, dass nur Instanzen mit Ereignissen, die im Ereignisstrom direkt benachbart sind, in die Ergebnismenge aufgenommen werden. Die Auswahlstrategien „skip-till-next-match“ und „robust-skip-till-next-match“ erlauben nur Instanzen mit den frühesten Ereignissen im Ereignisstrom nach dem Startereignis der Instanz.

In unserer Forschung haben wir eine Auswahlstrategie definiert, die die besten k (engl. top k) Instanzen erlaubt, die eine Scoring-Funktion über die Ereignisse einer Instanz maximieren. Eine solche Auswahlstrategie erlaubt es die Ergebnismenge anhand des Inhalts der Ereignisse in einer Instanz statt anhand der Position der Ereignisse im Ereignisstrom zu begrenzen. Zum Beispiel, um die wahrscheinlichste Instanz in einem Strom mit unsicheren Ereignissen zu finden, ist die zu maximierende Scoring-Funktion das Produkt der Auftrittswahrscheinlichkeiten der Ereignisse in einer Instanz. Ein weiteres Beispiel ist das Auffinden der Instanzen mit den meisten Ereignissen für ein Muster, das eine variable Anzahl von Ereignissen spezifiziert (z.B. durch eine Kleene-Hülle). In einem solchen Fall wird die Summe der Ereignisse einer Instanz maximiert.

An der top k Auswahlstrategie für EPM habe ich zusammen mit Panagiotis Bouros, der auch Postdoktorand bei METRIK war, gearbeitet. Wir haben die Auswahlstrategie definiert und einen Algorithmus entwickelt, der das Problem effizient löst. Eine vollständige Evaluierung des Algorithmus und die Publikation der Forschungsergebnisse konnten wir vor unserem Ausscheiden aus METRIK nicht umsetzen. Wir planen diese offenen Punkte in Zukunft wieder aufzugreifen und abzuschließen.

Die gesuchten zusammengesetzten Ereignisse anhand eines Musters zu beschreiben ist im Allgemeinen nicht trivial, weil es nicht immer klar ist, aus welchen Ereignissen sie tatsächlich bestehen und welche Bedingungen diese Ereignisse erfüllen müssen. Zum Beispiel wenn man mit Sensormessungen entlang einer Straße vorbeifahrende PKWs und LKWs zählen will, ist es nicht trivial ein Muster für das zusammengesetzte Ereignis "vorbeifahrender PKW" zu formulieren. Die Herausforderung ist, dass man a priori nicht weiß welche Sensormessungen wichtig für die Erkennung eines vorbeifahrenden PKWs sind.

Das Erlernen von Ereignismustern (engl. Event Pattern Learning - EPL) ist eine Technik des überwachten maschinellen Lernens, um Muster aus einer Menge von historischen Ereignissequenzen zu lernen. Die Menge der historischen Ereignissequenzen besteht aus positiven und negativen Sequenzen. Positive Sequenzen haben zum gesuchten zusammengesetzten Ereignis und negative Sequenzen haben nicht zum gesuchten zusammengesetzten Ereignis geführt. Die durch EPL erhaltenen Muster kann man in weiterer Folge in EPM Algorithmen verwenden, um die gesuchten zusammengesetzten Ereignisse in zukünftigen Ereignisströmen zu

finden. Der Vorteil von EPL gegenüber anderer Techniken des überwachten maschinellen Lernens, wie z.B. Support Vector Machines, ist, dass man schon vorhandene EPM Algorithmen zur Erkennung der zusammengesetzten Ereignisse verwenden kann und dass das erlernte Modell (Muster) explizit und dadurch interpretierbar ist.

Eine Beispielanwendung für EPL ist das Erlernen des Musters für die zusammengesetzten Ereignisse "vorbeifahrender PKW" und "vorbeifahrender LKW" aus historischen Ereignissequenzen. Dabei besteht die Menge der Ereignissequenzen für das Erlernen des zusammengesetzten Ereignisses "vorbeifahrender PKW" aus Sequenzen, bei denen ein PKW vorbeigefahren ist und aus Sequenzen bei denen kein PKW vorbeigefahren ist. Entsprechendes gilt für das Erlernen des zusammengesetzten Ereignisses "vorbeifahrender LKW". Diese Muster werden dann verwendet, um vorbeifahrende PKWs und LKWs mit Hilfe von EPM zu erkennen. Ziel dieser Forschung ist die Entwicklung effizienter Datenstrukturen und Algorithmen für EPL. Die Herausforderungen liegen in der Maximierung des Recalls und der Precision, sowie in der Minimierung der Ausführungszeit und des Speicherbedarfs.

An EPL habe ich zusammen mit METRIK-Doktorand Lars George geforscht. Das Problem wurde definiert und ein entsprechender Algorithmus wurde entwickelt. Die Evaluierungsergebnisse liegen vor und eine Publikation wird gerade vorbereitet. Die Publikation soll bei einer international renommierten Data Mining- oder Datenverarbeitungskonferenz (z.B. KDD, ICDM, ICDE, VLDB, SIGMOD) eingereicht werden. Für die Zukunft planen wir den Algorithmus hinsichtlich der Ausführungszeit und des Speicherbedarfs sowie der Ausdrucksstärke der erlernten Muster (z.B. Erlernen von Negationen und Kleene-Hüllen) zu verbessern.

Der dritte Bereich mit dem ich mich, neben EPM und EPL, beschäftigt habe, ist die verteilte und parallele Datenstromverarbeitung. Dabei beteiligte ich mich als Co-Autor am Forschungsvorhaben des METRIK-Doktoranden Matthias Sax. Ziel der Forschung ist es, den Ressourcenverbrauch (Anzahl der CPUs) Map/Reduce-basierter Datenstromverarbeitungssysteme (z. B. Apache Strom, Apache Flink) zu minimieren. Map/Reduce- basierter Datenstromverarbeitungssysteme erlauben Datenströme in Form von Datenflussprogrammen zu verarbeiten. Der Fokus bei der Verarbeitung von Datenflussprogrammen mit Datenströmen ist ein anderer als mit Datenmengen (engl. Batches). Während man bei der Batch-Verarbeitung eine endliche Datenmenge in möglichst kurzer Zeit verarbeiten will, will man bei der Datenstromverarbeitung einen potenziell unendlichen Datenstrom mit einem Durchsatz (verarbeitete Datenelemente pro Sekunde) verarbeiten, der mindestens so hoch ist wie die Geschwindigkeit des Eingangsdatenstroms (eintreffende Datenelemente pro Sekunde). Der Durchsatz soll mindestens so hoch wie die Geschwindigkeit des Eingangsdatenstroms sein, damit die Verarbeitung keinen Flaschenhals darstellt, der den Datenstrom verlangsamt. Dabei sollen so wenige

Ressourcen wie möglich verbraucht werden (Minimierung des Parallelisierungsgrades). Für die Berechnung des Durchsatzes eines Datenflussprogramms haben wir ein Kostenmodell definiert. Um den minimalen Parallelisierungsgrad zu berechnen, haben wir einen Algorithmus entwickelt, der anhand des Kostenmodells für jeden Operator im Datenflussprogramm den minimalen Parallelisierungsgrad bestimmt, sodass der minimale Durchsatz während der Verarbeitung nicht unterschritten wird.

Wir haben das Kostenmodell und den Algorithmus entwickelt und implementiert. Aktuell arbeiten wir an der Evaluierung und der Publikation unserer Forschungsergebnisse. Die Publikation soll bei einer renommierten Datenverarbeitungskonferenz (z.B. VLDB, SIGMOD, ICDE) eingereicht werden. Für die Zukunft planen wir einen Algorithmus zu entwickeln, der die Verteilung der Operatoren des Datenflussprogramms auf die CPUs mit der Minimierung des Parallelisierungsgrads verbindet. Dadurch werden zusätzliche Optimierungen möglich.

Forschungsaufenthalte, Konferenzen, Workshops:
Gemeinsamer Workshop der Informatik-Graduiertenkollegs Deutschlands, Schloss Dagstuhl, Juni 2014.

39th International Conference on Very Large Data Bases (VLDB), Riva del Garda, Italien, August 2013.

16. Fachtagung "Datenbanksysteme für Business, Technologie und Web" (BTW), Hamburg, März 2015.

Eigene Publikationen:
B. CADONNA, J. GAMPER, M. H. BÖHLEN: *A Robust Skip-Till-Next-Match Selection Strategy for Event Pattern Matching.* Proceedings of the 18th East-European Conference on Advances in Databases and Information Systems (ADBIS-14), Ohrid, Republic of Macedonia, September 2014.

PANAGIOTIS BOUROS

Analyzing Complex Data Types and Routing Problems

With the proliferation of mobile location-aware devices (e.g., smartphones) real-life data objects can be routinely “tagged” with different types of auxiliary information, such as text, spatial locations and time. For instance, photos on Flickr are assigned keywords and a spatial location. Current trend for services based upon the social interactions of their users gave rise to the so-called Location-Aware Social Networks (LASN), like Foursquare or Twitter (geo-tagged tweets). Based on their profile and activities, the users of such networks carry explicit or implicit spatial and textual information (e.g., posts, tweets) while associated at the same time with connectivity information derived from the social graph. Recently, the Resource Description Framework (RDF) used in the context of Semantic Web to define knowledge bases

has been extended to represent geographic information and support spatial queries. Finally, trajectory data capture the traveling history of moving objects such as people or vehicles. Under this, applications such as route recommendation, traveling behavior mining and disaster management call for efficient trajectory retrieval.

It is hence clear that data are nowadays becoming increasingly more complex. In the past, space, text, time and graphs have been well studied but in most cases independently. Recently, however, there has been a growing interest to investigate the correlation of these data dimensions, and for defining novel analysis tasks. In this context, disaster management provides an interesting case study as huge volumes of data are posted in real time on social media apart from the traditional news media, during catastrophic events. Twitter became a go-to platform during Hurricane Sandy hit in 2012; more than 20 million tweets were posted in a five day period. Geo-tagged tweets provided crucial information for both people and authorities in their effort to deal with the approach and aftermath of the hurricane.

In line with this trend, my research has focused on introducing novel analysis tasks and on designing efficient evaluation methods. Specifically, in [4], I worked on a generalized top-k join operator which operates on complex typed join attributes and predicates, e.g., spatial distance or string joins. Top-k joins have been extensively studied in relational databases, for the case where the join predicate is equality. The state-of-the-art evaluation algorithms aim at minimizing the number of accesses. However, when collections of complex data types are joined, computational cost can easily become the bottleneck. In view of this, we propose a novel evaluation paradigm, which minimizes the computational cost, without compromising the access cost. Further, in [5], I investigated the efficient management of spatial RDF data. Despite the work on efficient storing and querying knowledge bases, very little work exists on effective handling of spatial semantics in RDF. Our contributions include an effective encoding scheme for entities having spatial locations, the introduction of on-the-fly spatial filters and spatial join algorithms, and several optimizations that minimize the overhead of geometry and dictionary accesses. Finally, motivated from word-of-mouth and viral marketing campaigns, I also considered mining tasks on LASNs such as identifying the most influential users with respect to a spatial region [6].

Under a different perspective, the context of dealing with disaster events like earthquakes and other kind of natural catastrophes provides a good case study also for novel routing problems that go beyond the computation of the shortest or the fastest route. In [7], I studied the problem of providing meaningful routing directions. In this case, the fastest route may not be the ideal choice. Specifically, dealing with the aftermath of a natural disaster requires an evacuation plan to be communicated to people on site. Under such circumstances of distress and

disorganization, it is often desirable to provide concise, easy to memorize, and clear to follow instructions, and thus, the simplest route may be more preferable than the fastest route. Further, in [2], I studied the task of alternative routing. Shortest path computation is a fundamental problem on road networks which finds application in various domains of the research and the industry. Returning however solely this path is often not satisfying. Users seek alternative paths to enjoy more routing options and freedom. For instance, most commercial navigation systems recommend apart from the shortest path, a number of alternatives with different characteristics, leaving the final decision to the user. Given two nodes of the network, the goal is to recommend as set of k paths, which are sufficiently dissimilar to each other and as short as possible.

Recently in [3], I also investigated distance-based trajectory search; given a collection of trajectories and a set query points, the goal is to retrieve the top-k trajectories that pass as close as possible to all query points. Our work advanced the state-of-the-art by combining existing approaches to a hybrid method and also proposing an alternative, more efficient range-based approach. In addition, we proposed and studied the practical variant of bounded distance-based search, which takes into account the temporal characteristics of the searched trajectories.

Finally, I pursued the efficient computation of operators on complex data types, in the context of big data analysis. Specially, in [1], we focus on the evaluation of textual similarity joins by conducting a thorough experimental analysis. We compare the state-of-the-art methods particularly tailored for textual joins against general join methods. Our focus is to unveil the limits of the proposed methods, and investigate how the characteristics of the input datasets and the setup of the underlying MapReduce framework (e.g., Handoop) affects their performance.

Forschungsaufenthalte, Konferenzen, Workshops:
Research stay at The University of Hong Kong, China, May 4 - Jun 3, 2014.

Research stay at Free University of Bozen-Bolzano, Italy, October 22-25, 2013.

Dagstuhl Joint Workshop of the DFG Research Training Groups in Computer Science, May 27-29, 2013.

12th Hellenic Data Management Symposium (HDMS), Athens, Greece, July 24-25, 2014.

21st ACM International Conference on Advances in Geographic Information Systems (SIGSPATIAL GIS), Orlando FL, USA, November 5-8, 2013.

39th International Conference on Very Large Data Bases (VLDB), Riva del Garda, Trento, Italy, August 26-30, 2013.

13th International Symposium on Spatial and Temporal Databases (SSTD), Munich, Germany, August 21-23, 2013.

Eigene Publikationen:

[1] F. FIER AND P. BOUROS, *Textual Joins on MapReduce: An Experimental Evaluation,* 2015. (under preparation)

[2] T. CHONDROGIANNIS, P. BOUROS, J. GAMPER AND U. LESER, *Alternative Routing: k-Shortest Paths with Limited Overlap,* 2015 (under review)

[3] S. QI, P. BOUROS, D. SACHARIDIS AND N. MAMOULIS, *Efficient Point-based Trajectory Search,* in Proceedings of the 14th International Symposium on Spatial and Temporal Databases (SSTD), Hong Kong SAR, China, August 26-28, 2015 (to appear)

[4] S. QI, P. BOUROS AND N. MAMOULIS, *Efficient Top-k Join Processing on Complex Data Types,* 2014. (under review)

[5] J. LIAGOURIS, N. MAMOULIS, P. BOUROS AND M. TERROVITIS, *An Effective Encoding Scheme for Spatial RDF Data,* in Proceedings of the VLDB Endowment (PVLDB), Vol 7, No 12, 2014.

[6] P. BOUROS, D. SACHARIDIS AND N. BIKAKIS, *Regionally Influential Users in Location-Aware Social Networks,* in Proceedings of the 22nd ACM International Conference on Advances in Geographic Information Systems (SIGSPATIAL GIS), Dallas TX, USA, November 4-7, 2014.

[7] D. SACHARIDIS AND P. BOUROS, *Routing Directions: Keeping it Fast and Simple,* in Proceedings os the 21st ACM International Conference on Advances in Geographic Information Systems (SIGSPATIAL GIS), Orlando FL, USA, November 5-8, 2013.