Informatik Ruhr: Doktorandenkolleg 2011

Gernot A. Fink, Jan Vahrenhold

6./7. Oktober 2011

haus nordhelle

Valbert

Inhaltsverzeichnis

Entwurf stabiler DNA-Nanostrukturen

Dipl.-Inf. Mariannna D'Addario
Informatik, TU Dortmund
marianna.daddario@tu-dortmund.de
Prof. Dr. Sven Rahmann

PROBLEMBESCHREIBUNG UND FORSCHUNGSFRAGE

Das Ziel dieser Arbeit ist der Entwurf von stabilen DNA-Nanostrukturen. DNA-Nanostrukturen haben eine Größe von wenigen Nanometern (1nm $= 1 \times 10^{-9}$m) und bestehen aus mehreren DNA-Sequenzen. Eine DNA-Sequenz ist ein String über dem DNA-Alphabet $\{A, C, G, T\}$. DNA-Nanostrukturen sind stabil, wenn die DNA-Sequenzen nur an vordefinierten Stellen hybridisieren und eine gewünschte Form einnehmen. Zwei Sequenzen hybridisieren miteinander und bilden die bekannte DNA-Doppelhelix, wenn diese Watson-Crick-komplementär zueinander sind. Das Watson-Crick-Komplement alleine ist allerdings kein Garant dafür, dass vordefinierte DNA-Nanostrukturen entstehen. Zunächst müssen beim Sequenzdesign Kreuzhybridisierungen vermieden werden. Alle beteiligten Sequenzen sollten so entworfen werden, dass sie nur an eindeutigen Stellen hybridisieren und keine weiteren Möglichkeiten haben Bindungen einzugehen. Ein weiteres Problem sind die thermodynamischen Eigenschaften der Strukturen. Die Einflüsse dieser Eigenschaften auf DNA-Nanostrukturen sind bisher nicht im Detail erforscht worden. Kriterien zum Entwurf stabiler DNA-Nanostrukturen würden der DNA-Nanotechnologie ermöglichen Bottom-Up-Verfahren einzusetzen. Diese Kriterien müssten dabei sowohl das Sequenzdesign als auch die thermodynamischen Eigenschaften vereinen.

VORGEHENSWEISE UND METHODE

Zunächst werden theoretisch geeignete Kriterien für den Entwurf einer Stuktur gewählt. Unter Berücksichtigung dieser Kriterien werden dann einige DNA-Strukturen entworfen. Anschließend findet eine experimentelle Verifikation im Labor statt. Die entworfenen DNA-Strukturen werden assembliert und durch Gelelektrophorese evaluiert. Die Laborergebnisse gehen dann als Grundlage zur Entwicklung weiterer Kriterien oder zum Ausschluss vorheriger Kriterien ein. Schon bekannte Methoden wie DNA-Origami [3], zum Entwurf von DNA-Strukturen einer vordefinierten Form, beziehen sich auf größere Strukturen (etwa 100 nm). Die DNA-Origami bestehen aus hunderten DNA-Sequenzen und stehen weniger unter dem Einfluss thermodynamischer Eigenschaften. In Origami-Strukturen sind begrenzte Unstimmigkeiten einzelner Watson-Crick-Komplemente tolerierbar, da das System als Ganzes nicht davon betroffen ist. Diese Fehlertoleranz ist in kleinen DNA-Nanostrukturen, die aus ca. zehn Sequenzen bestehen, nicht realisierbar. In diesen Fällen entsteht die DNA-Nanostruktur nicht in der geforderten Form, sondern es bilden sich Teilprodukte oder unerwünschte Sekundärstrukturen.

VERWANDTE ARBEITEN

Die Arbeit von Seeman [5] hat den Grundstein zur DNA-Nanotechnologie gelegt. Das sequenzdesigntheorethische Problemen vordefinierter DNA-Strukturen wurde in der Arbeit von Udo Feldkamp [2] definiert. Christof Niemeyer und Udo Feldkamp beschreiben in ihrer Arbeit [1] das übergeordnete Ziel von DNA-Nanoarchitekturen. Der Ansatz zum Entwurf der DNA-Sequenzen von Udo Feldkamp nutzt einen Graphen, in welchem die Knoten wichtige Informationen zur Bildung der Sequenzen enthalten. Ein anderer Ansatz ist die Kanten eines De-Bruijn-Graphen als Informationsträger zu nutzen. Mit dieser Modifikation kann das Sequenzdesign Problem auf das Eulerkreis Problem reduziert werden. In den genannten Arbeiten wurden allerdings keine thermodynamischen Kriterien zum Entwurf von DNA-Sequenzen berücksichtigt. Diese sollen zunächst identifiziert werden und anschließend in den Entwurf einfließen.

VORLÄUFIGE ERGEBNISSE

Ein Tool zum Entwurf von Sequenzen für eine bestimmte DNA-Nanostruktur, die 4×4-Kachel [4], wurde implementiert und mit verschiedenen Analysefunktionen erweitert. Im Juni 2011 wurden erste Kacheln ausgewählt und im Labor synthetisiert. Die theoretisch erarbeiteten Kriterien waren: die Reihenfolge der Hybridisierungen, die gesamte Temperaturspanne der Assemblierung und die Temperaturspannen der einzelnen Domänen der DNA-Struktur. Die Laborergebnisse haben das Kriterium bezüglich der Reihenfolge der Hybridisierungen verifiziert. Die DNA-Struktur formt sich am besten, wenn die Reihenfolge strikt eingehalten wird. Dieses und zwei weitere thermodynamische Kriterien sind beim Entwurf der nächsten Strukturen berücksichtigt worden. Zwei der entworfenen Strukturen wurden im August 2011 ausgewählt, um in der bevorstehenden experimentellen Verifikation (September 2011) synthetisiert zu werden. Die Versuche werden in Eigenarbeit im Labor durchgeführt.

WEITERE SCHRITTE

Einzelne DNA-Nanostrukturen können als Bausteine für skalierbare DNA-Architekturen genutzt werden. Zwei Kacheln können in Abhängigkeit voneinander so entworfen werden, dass sie abwechselnd miteinander hybridisieren. Existieren beispielsweise die Kachel „Weiss" und die Kachel „Schwarz", dann entsteht ein Schachbrett beliebiger Größe. Um eine solche Architektur zu ermöglichen, müssen die einzelnen DNA-Kacheln stabil sein. Eine andere Möglichkeit die Stabilität zu erhöhen könnte die Änderung der

gewünschten Form der DNA-Strukturen sein. Statt kreuzartiger Formen wären auch dreiarmige Strukturen denkbar. Die Kooperation mit der Arbeitsgruppe „Chemische Biologie“ von Prof. Dr. C. M. Niemeyer, Fakultät Chemie, TU Dortmund, ist für diese Arbeit unverzichtbar. Bisher wurden alle Ergebnisse und Laborversuche in Zusammenarbeit mit dieser Arbeitsgruppe erarbeitet und durchgeführt.

REFERENCES

1. FELDKAMP, U. und C.M. NIEMEYER: *Rationaler Entwurf von DNA-Nanoarchitekturen*. Angewandte Chemie, 118(12):1888–1910, 2006.
2. FELDKAMP, U., H. RAUHE und W. BANZHAF: *Software tools for DNA sequence design*. Genetic Programming and Evolvable Machines, 4(2):153–171, 2003.
3. ROTHEMUND, PW: *Folding DNA to create nanoscale shapes and patterns.-pp 297-302*. Nature, 440(7082), 2006.
4. SACCÀ, B., R. MEYER und C.M. NIEMEYER: *Analysis of the Self-Assembly of 4× 4 DNA Tiles by Temperature-Dependent FRET Spectroscopy*. ChemPhysChem, 10(18):3239–3248, 2009.
5. SEEMAN, N.C. und N.R. KALLENBACH: *Design of immobile nucleic acid junctions*. Biophysical journal, 44(2):201–209, 1983.

Rational reasoning with conditionals and probabilities

Christian Eichhorn
Fakultät für Informatik, TU Dortmund
christian.eichhorn@tu-dortmund.de
Prof. Dr. Gabriele Kern-Isberner

PROBLEMBESCHREIBUNG UND FORSCHUNGSFRAGE

Die Arbeit ist Teil eines Projektes des Schwerpunktprogramms (SPP) 1516 der deutschen Forschungsgesellschaft (DFG), in welchem untersucht wird, in wieweit die normativen Theorien der Rationalität mit der aktuellen empirischen Forschung in der Psychologie übereinstimmen [5]).

Häufig entscheiden Menschen auf Basis von konditionalem Wissen [3], zum Beispiel, dass, sollte man für die anstehende Prüfung nicht intensiver lernen, man durchfallen könnte (als Konditional *(durchfallen|nicht_lernen)*).

Im Rahmen dieser Arbeit werden Methoden, die auf konditionalem Wissen aufsetzen, insbesondere *c-Repräsentationen* [4], im Hinblick auf ihren Einfluss auf menschliches Schlussfolgern untersucht:

- Sind Logiken auf Basis konditionaler Stukturen ein adäquates Modell für menschliches Schlussfolgern?
- Können konditionale Strukturen als „Strukturen rationalen Denkens“ verstanden werden?
- Welche c-Repräsentation ist die „beste“?
- In welcher Beziehung stehen c-Repräsentationen zur Frage nach Kohärenz von konditionalen Wissensbasen [2]?

Mit dem auf diesen Erkenntnissen aufbauenden Kernthema:

- Entwicklung von Methoden und Algorithmen zum Einsatz in Systeme und intelligenten Agenten.

Für den Bereich der *künstlichen Intelligenz* soll hierdurch die Möglichkeiten verbessert werden, menschliches Schlussfolgern nachzuempfinden, zu antizipieren oder zu simulieren.

Für den Bereich der *Psychologie* und *Philosophie* werden aufgrund dieser Erkenntnisse bessere Erklärungsmodelle dafür erwartet, wie Menschen auf Basis von konditionalem Wissen „vernünftig“ schlussfolgern (Vgl. [3]).

VORGEHENSWEISE UND METHODE

Für den Bereich der Unterschungen von konditionalen Strukturen, c-Repräsentation, Kohärenz und weiteren bestehenden Methoden wie unter anderem Schlussfolgern mittels maximaler Entropie [1] und System Z [6] wird die Forschung theoretischer Natur sein.

Auf Basis dieser Erkenntnisse wird untersucht, ob die entwickelten Modelle menschliches Schlussfolgern adäquat abbilden. Dies wird innerhalb des Projektes durch psychologische Experimente (Vgl. [7]) geprüft. Die Mitentwicklung der Fragebögen und die Auswertung der Testergebnisse werden Teil der Arbeit sein.

VERWANDTE ARBEITEN

Die Arbeit baut auf auf dem Kohärenzansatz von Gilio [2], den Arbeiten von Spohn zu Konditionalen [8], den Arbeiten zu konditionalen Wissensbasen von Kern-Isberner [4], sowie den Untersuchungen von Pfeifer und Kleiter [7] und sucht Verbindungen zwischen ebendiesen.

VORLÄUFIGE ERGEBNISSE & WEITERE SCHRITTE

Da sich die Arbeit noch im Stadium der Vorbereitung befindet, können weder vorläufige Ergebnisse noch weitere Schritte angegeben werden.

REFERENCES

1. C. Beierle, G. Kern-Isberner. *Methoden wissensbasierter Systeme*. 4. Auflage. Vieweg + Teubner. 2008.
2. A. Gilio. *Probabilistic Reasoning Under Coherence in System P*. Annals of Mathematics and Artificial Intelligence, 34(1-3): S. 5–34. 2002.
3. G. Kern-Isberner, N. Pfeifer. *Rational reasoning with conditionals and probabilities*. Projektantrag zum SPP 1516. 2010. Unveröffentlicht.
4. G. Kern-Isberner. *Conditionals in Nonmonotonic Reasoning and Belief Revision*. Lectures Notes in Computer Science, 2087. Springer. 2001.
5. M. Knauff, R. Hertwig, G. Schurz, W. Spohn, M. Waldmann. *Proposal for the Establishment of a Priority Program – New Frameworks of Rationality*. Online: Proposal for the Establishment of a Priority Program 2010.
6. J. Pearl. *System Z: a natural ordering of defaults with tractable applications to default reasoning*. In Proceedings of TARK: S. 121–135. 1990.
7. N. Pfeifer, G. Kleiter. *Coherence and Nonmonotonicity in Human Reasoning*. Synthese, 146: S. 93–109. Springer. 2005.
8. W. Spohn. *Ordinal Conditional Functions: A Dynamic Theory of Epistemic States*, in *Causation, Coherence, and Concepts*. Boston Studies in the Philosophy of Science, 256: 19–41. Springer. 2008.

On the Value of Social Web Data in Recommender Systems

Fatih Gedikli
Fachbereich Informatik, LS XIII, TU Dortmund
fatih.gedikli@tu-dortmund.de
Prof. Dr. Dietmar Jannach

PROBLEMBESCHREIBUNG UND FORSCHUNGSFRAGE

Empfehlungssysteme sind Systeme, die dem Benutzer automatisch personalisierte Empfehlungen generieren. Empfehlungssysteme können nach verschiedenen Gesichtspunkten evaluiert werden, wie z.B. nach der Transparenz oder der Genauigkeit der gemachten Empfehlungen.

Das Ziel meiner Forschungstätigkeit besteht darin, durch Ausnutzung von Daten aus dem sog. Social Web, die Qualität der Empfehlungen von Empfehlungssystemen zu verbessern. Eine höhere Qualität der Empfehlungen führt zu zufriedeneren Kunden, was wiederum mehr Umsatz für ein Unternehmen bedeutet, das Empfehlungssysteme einsetzt.

In meiner Arbeit möchte ich herausfinden, in wie weit ich Daten aus dem Social Web ausnutzen kann, um unterschiedliche Aspekte von Empfehlungssystemen zu verbessern. Transparenz, Genauigkeit, Effizienz und Überredungsfähigkeit sind Beispiele für solche Kriterien, die in meiner Arbeit untersucht werden.

VORGEHENSWEISE UND METHODE

Bei der Evaluierungsmethodik richten wir uns nach den Vorgehensweisen in den verwandten Arbeiten auf dem Gebiet der Empfehlungssyteme. Durch Anlehnung an international anerkannte und viel zitierte Arbeiten versuchen wir die Qualität und Richtigkeit unserer Ergebnisse sicherzustellen.

In der Regel führen wir Offline-Experimente durch. Wir benutzen dabei Datensätze und Evaluierungs-Metriken, die auch andere Forscher in ihren Experimenten einsetzen. So können Algorithmen für Empfehlungssysteme bspw. durch eine Kreuzvalidierung nach ihrer Genauigkeit hin evaluiert werden. Wir führen aber auch Benutzerstudien durch, um auch Aspekte, die man in Offline-Experimenten schlecht messen kann, zu analysieren.

ERGEBNISSE

In unseren Forschungstätigkeiten konnten wir in unterschiedlichen Bereichen neue Ergebnisse gewinnen:

- Zum einen haben wir einen neuen Empfehlungsalgorithmus mit dem Namen *RF-Rec* vorgestellt, der effizienter und meistens sogar genauer als die aktuellen Vergleichsalgorithmen (wie z.B. der Algorithmus von Koren [2]) arbeitet.
- Zum anderen haben wir einen neuen Empfehlungsalgorithmus für Tags vorgestellt (*LocalRank*), der ebenfalls effizienter als der aktuelle Vergleichsalgorithmus *FolkRank* [1] arbeitet und zudem gleich gute Empfehlungen generiert.
- Wir haben das Konzept der produkt-gebundenen Tag-Präferenzen (*item-specific tag preferences*) eingeführt, mit der Benutzer Produkte präziser bewerten können. Diese Arbeit basiert auf der Idee von Sen et al. [3].
- Wir haben zudem gezeigt, dass wir durch die Ausnutzen der zusätzlichen Tag-Präferenzen genauere Vorhersagen berechnen können. Zudem haben wir gezeigt, wie man Tag-Präferenzen in Erklärungen für Empfehlungssysteme einsetzen kann.

Derzeit befinde ich mich in der Endphase meiner Promotion, in der ich die bereits publizierten Ergebnisse nur noch zusammenschreiben und abgeben muss.

REFERENCES

1. Hotho, A., Jäschke, R., Schmitz, C., Stumme, G.: Information retrieval in folksonomies: Search and ranking. In: Proc. ESWC'2006, Budva, ME (2006) 411-426.

2. Y. Koren: Factor in the neighbors: Scalable and accurate collaborative filtering, ACM Transactions on Knowledge Discovery from Data, vol. 4, pp. 1:1–1:24, January 2010.

3. Sen, S., Vig, J., and Riedl, J:. Tagommenders: Connecting users to items through tags. In Proceedings of the 18th International World Wide Web Conference (WWW'09). Madrid, Spain, 671-680.

Geometriekalibrierung verteilter Sensorfelder

Marius Hennecke
LS12, Informatik, TU Dortmund
Marius.Hennecke@tu-dortmund.de
Prof. Dr.-Ing. Gernot A. Fink

PROBLEMBESCHREIBUNG UND FORSCHUNGSFRAGE

Sensorfelder sind eine wichtige Informationsquelle von Kontextinformationen für ein maschinelles Perzeptionssystem. So können Personen beispielsweise durch räumlich verteilte passive Infrarotstrahlungssensoren anhand ihrer Körperwärme lokalisiert werden. Akustische Ereignisse hingegen können mit Hilfe von verteilten Mikrofonfeldern lokalisiert werden. Nahezu alle Verfahren zur räumlichen Kombination der Daten individueller Sensoren setzen eine bekannte Sensorfeldgeometrie voraus.

Für den praktischen Einsatz verteilter Sensorfelder ist eine automatische Geometriekalibrierung unerlässlich. Insbesondere der Einsatz von Ad-Hoc Sensorfeldern, beispielsweise die Verwendung von mehreren Mobiltelefonen als Mikrofonfeld, macht eine solche Methode notwendig. Das zentrale Forschungsvorhaben ist die Entwicklung von Verfahren zur automatischen Geometriekalibrierung unterschiedlicher Sensorfelder. Mit Blick auf die Kalibrierung von Ad-Hoc Sensorfeldern sind insbesondere Verfahren interessant, die eine Änderung der Geometrie feststellen und die Kalibrierung bei Bedarf nachführen können. Darüber hinaus sollen Methoden zur Fusion verschiedener Sensor-Koordinatensysteme entwickelt werden.

VORGEHENSWEISE UND METHODE

Methoden zur automatischen Geometriekalibrierung können zunächst in überwachte und unüberwachte Verfahren eingeteilt werden. Eine überwachte Kalibrierung verwendet bekannte Signale von ggf. bekannten Positionen, die es beispielsweise erlauben Distanzen zwischen Sensorpaaren zu schätzen. Eine unüberwachte Kalibrierung hingegen nutzt lediglich die natürlich vorhandenen Signale. Für akustische Sensornetze sind das Umgebungsrauschen, Sprache sowie andere Geräusche. In einem unüberwachten Szenario wird die Kalibrierung meist durch die Tatsache erschwert, dass lediglich Lokalisierungsinformationen in Form von Raumwinkeln zur Verfügung stehen.

Aus Winkelmessungen für räumlich diversitäre Ereignisse ist die Inferenz aller Mikrofonkoordinaten möglich. Dazu wird die Kalibrierung als Optimierungsproblem formuliert, welches eine geeignete Modellbewertungsfunktion minimiert. Die zu optimierenden Parameter umfassen die Sensorfeldkoordinaten und die Sensorfeldorientierungen, die in Abhängigkeit der ebenfalls unbekannten absoluten Positionen der Kalibrier-Ereignisse bestimmt werden müssen.

VERWANDTE ARBEITEN

Die meisten Arbeiten zur Geometriekalibrierung von Sensorfeldern gehen von einem überwachten Szenario aus, in dem entweder die Sensoren selbst zusätzlich über einen Sender verfügen oder aber die Kalibrierungssignale von bekannten Positionen gesendet werden. In [3] werden beispielsweise fünf pyramidenförmig angeordnete Lautsprecher genutzt um die 448 Mikrofone des *Huge Microphone Arrays* zu kalibrieren. Es finden sich in der Literatur wenige Arbeiten zur vollständig unüberwachten Kalibrierung.

VORLÄUFIGE ERGEBNISSE

Erste Arbeiten zeigen die Möglichkeit einer unüberwachten Kalibrierung zweidimensionaler Feldgeometrien [1, 4]. Es wurde ebenfalls ein überwachtes Verfahren für Ad-Hoc Mikrofonfelder bestehend aus Mobiltelefonen entwickelt [2].

WEITERE SCHRITTE

Die Beseitigung der Einschränkungen bekannter Verfahren hin zu einer vollständig unüberwachten Geometriekalibrierung ist ein wichtiger Beitrag, um den praktischen Einsatz von Sensorfeldern in Zukunft zu ermöglichen. Hierbei ist insbesondere die Erweiterung auf dreidimensionale Feldgeometrien erforderlich. Ein weiterer wichtiger Schritt besteht in der automatischen Schätzung von Abbildungen unterschiedlicher Sensor-Koordinatensysteme aufeinander, so dass heterogene Sensoren zu einem gemeinsamen Sensorfeld fusioniert werden können.

REFERENCES

1. M. Hennecke, T. Plötz, G. A. Fink, J. Schmalenströer, and R. Häb-Umbach. A Hierarchical Approach to Unsupervised Shape Calibration of Microphone Array Networks. In *IEEE Workshop on Statistical Signal Processing*, Cardiff, UK, 2009.
2. M. H. Hennecke and G. A. Fink. Towards Acoustic Self-Localization of Ad Hoc Smartphone Arrays. In *Hands-Free Speech Communication and Microphone Arrays*, Edinburgh, UK, 2011.
3. J. M. Sachar, H. F. Silverman, and W. R. Patterson. Microphone position and gain calibration for a large-aperture microphone array. *IEEE Trans. Speech Audio Process.*, 13(1):42–52, 2005.
4. J. Schmalenstroer, F. Jacob, R. Haeb-Umbach, M. H. Hennecke, and G. A. Fink. Unsupervised Geometry Calibration of Acoustic Sensor Networks using Source Correspondences. In *Interspeech*, Florence, Italy, 2011.

Algorithms for the Investigation of Genotype and Phenotype

Johannes Köster
Informatik XI, TU Dortmund
johannes.koester@tu-dortmund.de
Prof. Dr. Sven Rahmann

PROBLEM

The genome contains the hereditary information of an organism. It consists of genes, that encode proteins that are built by the translational machinery. Individual organisms can be differentiated by their genome, or genotype. Even between two individuums of the same species the genotype differs. By encoding proteins, the genotype determines the phenotype of a cell to a large extent. In this scope, the phenotype is the entirety of physical and functional properties that are expressed by a cell. Rather than emerging directly from individual genes, these properties are generated by the cooperation of multiple proteins in large networks. On the one hand, these networks show complex regulation mechanisms among proteins, for example by allosteric regulation or competition on binding sites. On the other hand, they can regulate genes, allowing the cell to react on external signals by changing the expression of genes. Moreover, gene regulation again can have an effect on the regulatory network itself. Modern high throughput technologies allow large scale studies of both views. The thesis aims to investigate solutions to improve genotype and phenotype analysis in various ways.

Genotype analysis using high-throughput sequencing is still in its infancy. The thesis may address several weaknesses of current analysis pipelines. E.g. mapping RNA-reads using a GPGPU to provide increased performance, or a robust estimation of gene expressions that takes biases into account. In phenotype – i.e. protein network – analysis, predictive models suffer either from being to detailed to stay feasible for large-scale data (e.g. differential expression based models) or they are not able to capture the functional implications by regulatory mechanisms (e.g. graph based models). Based on my diploma thesis, a novel functionally predictive model for protein networks, called “protein hypernetworks” shall be improved and established. Ultimately genotype and phenotype analysis may be combined to provide a deeper insight into biological mechanisms.

ORGANIZATION AND METHODS

The algorithms and models developed in the thesis are mainly investigated in two cooperative projects. First, protein hypernetworks are used to analyse the human adhesome network together with the Max-Planck-Institute of Molecular Physiology Dortmund (Dr. Eli Zamir). Here, model-based predictions are verified in cooperations with system biologists using orthogonal experimental techniques like Fluorescence Microscopy. Second, genotype approaches are verified in cooperation with the University Hospital Essen (Dr. Alexander Schramm) and applied to the analysis of neuroblastoma cancer.

RELATED WORK

The protein hypernetworks model was initially developed in my diploma thesis. It was inspired by [3] who proved that considering mutually exclusive interactions can improve the quality of protein complex prediction. However, protein hypernetworks have evolved to be a much more general framework that allows to model all kinds of qualitative regulatory mechanisms and other predictions besides protein complexes. Mapping RNA-reads on the GPU was recently published by [1]. However, the used approach only works for smaller prokaryotic genomes. My plan is to explore alternative string matching techniques that allow for an improved performance (e.g. using bit-parallelism). Robust gene or transcript expression analysis considering biases was proposed e.g. by [2].

PRELIMINARY RESULTS AND NEXT STEPS

protein hypernetworks received major improvements since January 2011 (the beginning of my thesis). A thorough description of the framework and first biological applications onto the yeast protein network is now under review at “Bioinformatics”. The results show that protein hypernetworks can improve the quality of protein complex prediction and the prediction of functional importance of proteins. Importantly, we could show that the model allows to infer predictions about genetic interactions from the phenotypical network information. This is a first step to merge the two levels of analysis. The provided results where compared against gold standard data, e.g. the CYC2008 complex catalog and the SGD protein database.

Besides the application of protein hypernetworks on other organisms and networks like the human adhesome, I plan to implement novel predictions like functional similarity. Further, in cooperation with the MPI, the impact of evolution on the development of regulatory mechanisms will be investigated using the model.

REFERENCES

1. Jochen Blom, Tobias Jakobi, Daniel Doppmeier, Sebastian Jaenicke, Jörn Kalinowski, Jens Stoye, and Alexander Goesmann. Exact and complete short read alignment to microbial genomes using GPU programming. *Bioinformatics*, 27(10):1351–1358, March 2011.

2. Regina Bohnert and Gunnar Rätsch. rQuant.web: a tool for RNA-Seq-based transcript quantitation. *Nucleic Acids Research*, 38(suppl 2):W348–W351, July 2010.

3. Suk H. Jung, Bora Hyun, Woo-Hyuk Jang, Hee-Young Hur, and Dong-Soo Han. Protein complex prediction based on simultaneous protein interaction network. *Bioinformatics*, 26(3):385–391, 2010.

Ressourcen-beschränkte Analyse von Spektrometriedaten

Dipl.-Inf. Dominik Kopczynski
Fakultät für Informatik - TU Dortmund
dominik.kopczynski@tu-dortmund.de
Betreuer: Prof. Dr. Sven Rahmann

PROBLEMBESCHREIBUNG UND FORSCHUNGSFRAGE

Ionenmobilitätsspektrometrie (IMS) ist ein Verfahren zur Konzentrationsmessung von flüchtigen Stoffen in der Luft. Das MCC/IMS Spektrum ist ein zweidimensionales Spektrum mit der Retentions- und Driftzeit als Dimensionen, welches als Signal die Konzentration der im IMS-Gerät ionisierten Stoffe bzw. die somit erzeugte Spannung misst.
Durch eine Parametrisierung der in der Messung vorkommenden Peaks ist es möglich ein ganzes Spektrum, welches bei einer hochaufgelösten Messung ca. 300 Millionen Datenpunkte umfasst, mit nur wenigen Parametern pro Peak (7 Parameter) beschreiben zu können. Dies ist ein großer Vorteil bei der Entwicklung eines miniaturisierten IMS-Gerätes, da in einem eingebetteten System der Speicher immer einen hohen Energieverbrauch hat und somit die Nutzung des Speichers gering gehalten werden muss.

VORGEHENSWEISE UND METHODE

Die Idee ist, während der Messzeit nur eine geringe Anzahl der letzten gemessenen Driftzeit-Spektren zu speichern, um somit die Nutzung des Arbeitsspeichers möglichst gering zu halten. Die einzelnen Spektren werden vorverarbeitet und mit einem stochastischen Parametrisierungsverfahren zuerst auf die Anzahl der Peaks überprüft und anschließend parametrisiert. Aus den parametrisieren 1D Spektren werden anschließend die Parameter für die 2D Modelle der einzelnen Peaks bestimmt. Hierbei spielt der EM-Algorithmus[2] eine wichtige Rolle, welcher mit maximierenden Schätzern aller Parameter aus dem Modell ein Spektrum optimal beschreibt.

VERWANDTE ARBEITEN

PD. Dr. Jörg-Ingo Baumbach hat bereits mit seiner Forschungsgruppe auf dem Gebiet der MCC/IMS erhebliche Arbeit geleistet. U. A. beschreibt Vogtland[3] in seiner Dimplomarbeit Glättungs- und Fittingverfahren zur Detektion und Beschreibung von Peaks. Bader hat in ihrer Doktorarbeit[1] ein Verfahren zum Finden von Regionen in einem Spektrum vorgestellt. Jedoch besteht bei den Arbeiten das Problem, dass angenommen wird, dass ein Datenpunkt exakt zu einem Peak gehört, was bei sich überlappenden Peaks nicht der Fall ist. Zum Anderen ist das von Baumbach[4] vorgestellte 1D Modell eines IMS-Peaks in einem Spektrum eine Mischung aus Normal- und Breit-Wigner-Verteilung. Wünschenswert wäre nur eine Verteilung für einen Peak.

VORLÄUFIGE ERGEBNISSE

Eine Online-Parametrisierung (während der Messung) ist bereits etabliert und liefert mit dem neu entwichelten Modell akzeptable Ergebnisse. Es wird eine Liste aller Peaks erzeugt, welche in der Messung gefunden und parametrisiert werden konnten. Im nächsten Schritt müssen diese Peaks einheitlich aligniert werden, da jedes IMS-Gerät leichte Fluktuationen in Aufbau und Parametereinstellungen aufweist.

WEITERE SCHRITTE

- Vorhersage von Spektren: Ein wichtiger Punkt ist es mit Zuhilfenahme der physikalischen und chemischen Eigenschaften eines Stoffes eine Vorhersage zu treffen, zu welcher Retentions- und Driftzeit ein Peak entsteht. Dieses Problem stellt sich als besonders schwierig heraus, da die Berechnung mitunter eine Betrachtung der Substanzen auf molekularer Ebene in Betracht ziehen muss um die Interaktion zwischen den Molekülen und den Messeinheiten exakt zu bescheiben.

- Evaluierung von Analyseverfahren: In dieser Phase sollen alle Verfahren, die zur Analyse und Auswertung von IMS Daten eingesetzt werden (Vorverarbeitung, Peakerkennung, Parametrisierung, etc.) verglichen und evaluiert werden. Die Idee ist es ein Standardverfahren zu etablieren, auf das sich zukünftige Verfahren und Algorithmen beziehen können, um ihre Güte zu analysieren.

REFERENCES

1. BADER, S.: *Identification and Quantification of Peaks in Spectrometric Data*. Doktorarbeit, TU Dortmund, 2008.

2. BILMES, J.: *A Gentle Tutorial of the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models*. Technischer Bericht, 1998.

3. VOGTLAND, D.: *Untersuchung von Ionenmobilitätsspektrometriedaten auf Peaks von Substanzen in verschiedenen Konzentrationen unter Einsatz von Glättungs- und Fittingverfahren*. Diplomarbeit, TU Dortmund, 2007.

4. VOGTLAND, D. und J.I. BAUMBACH: *Breit-Wigner-Function and IMS-signals*. International Journal for Ion Mobility Spectrometry, 12:109–114, 2009. 10.1007/s12127-009-0027-8.

Algorithm Engineering für Probleme aus der Chemieinformatik

Nils M. Kriege
Fakultät für Informatik, TU Dortmund
nils.kriege@tu-dortmund.de
Prof. Dr. Petra Mutzel

PROBLEMBESCHREIBUNG UND FORSCHUNGSFRAGE

Die Chemie- und Bioinformatik beschäftigt sich mit Molekülen, die häufig als Graphen modelliert werden. Eine grundlegende Fragestellung, die als Teilproblem bei vielen Anwendungen wie der Suche in Graphdatenbanken oder bei Verfahren, die sich dem maschinellen Lernen zuordnen lassen, auftritt, besteht in dem automatisierten Vergleich von Graphen. Hierfür sind verschiedene graphentheoretische Probleme relevant, die sowohl theoretisch als auch praktisch gut untersucht sind. Hierzu zählen beispielsweise das (Sub-) Graph Isomorphie Problem so wie das Maximum Common Subgraph (MCS) Problem, für welche keine Polynomialzeit-Algorithmen für allgemeine Graphen bekannt sind.

In der Praxis müssen diese Probleme jedoch nicht zwangsläufig exakt gelöst werden. Zudem eröffnet das Ausnutzen spezieller Eigenschaften der betrachteten Graphen die Möglichkeit, exakte Lösungen effizient zu finden. Beispielsweise haben Molekülgraphen beschränkten Grad und sind fast immer planar. Zudem können sich aus der Anwendung Anforderungen an ein Ähnlichkeitsmaß ergeben: Die Ähnlichkeitssuche kann beispielsweise davon profitieren, dass ein definiertes Distanzmaß metrisch ist und die Verwendung im Rahmen von SVMs erfordert eine geeignete Kernel-Funktion, die die Ähnlichkeit von Graphen angibt. Eine konkrete Anwendung finden diese Techniken in der Pharmaforschung, bei der zunehmend computergestützte Methoden eingesetzt werden, um z.B. den Umfang kostspieliger experimenteller Untersuchungen zu reduzieren.

Im Rahmen der Arbeit sollen Methoden entwickelt werden, die diesen unterschiedlichen Anforderungen Rechnung tragen, indem einerseits komplexitätstheoretisch untersucht wird, unter welchen Umständen exakte Polynomialzeit-Algorithmen für Molekülgraphen möglich sind, und andererseits Verfahren für die Praxis entworfen und evaluiert werden.

VORGEHENSWEISE UND METHODE

Die Vorgehensweise orientiert sich an dem im *Algorithm Engineering* typischen Kreislauf und beinhaltet den Entwurf, die Analyse und experimentelle Untersuchung von Algorithmen, wobei experimentell gewonnene Erkenntnisse ggf. wiederum zur Anpassungen des Algorithmus oder seiner Implementierung führen können. Die praktische Untersuchung beinhaltet den Vergleich mit bekannten Verfahren auf Benchmarkinstanzen und berücksichtigt neben der Laufzeit ggf. auch die Angemessenheit des Ähnlichkeitsmaßes, z.B. durch einen Vergleich der Vorhersagegenauigkeit mit Hilfe von SVMs.

VERWANDTE ARBEITEN

Die Arbeit [2] stellt einen Algorithmus für das MCS-Problem vor, der sowohl graphentheoretische Zusammenhänge nutzt als auch praktisch relevante Heuristiken zur Beschleunigung einsetzt. In [3] wird ein MCS-Algorithmus mit polynomieller Laufzeit für außenplanare Graphen vorgeschlagen unter der Einschränkung, dass Blöcke und Brücken erhalten bleiben, was aus chemischer Sicht für Molekülgraphen sinnvoll ist. Die Arbeit [4] befasst sich mit dem Vergleich von Graphen mit Hilfe von Kernels.

Bis auf wenige Ausnahmen ist Theorie und Praxis weitgehend getrennt betrachtet worden. Die Auswirkungen von Nebenbedingungen aus der Praxis auf die Komplexität des zu Grunde liegenden Problems ist unzureichend untersucht und verspricht neue Erkenntnisse.

VORLÄUFIGE ERGEBNISSE

Für die Subgraph-Suche in Graphdatenbanken wurde ein effizientes Verfahren entwickelt, das mit Hilfe einer Indexstruktur die Suche heuristisch auf eine kleinere Kandidatenmenge beschränkt, die jedoch falsch-positive Instanzen enthalten kann. Mit Hilfe eines speziell für diese Anwendung entworfenen Subgraph-Isomorphie-Algorithmus wird die Kandidatenmenge daher anschließend überprüft [1].

Außerdem wurde ein Graph Kernel entwickelt, der auf Isomorphismen gemeinsamer Subgraphen beruht und einen flexiblen Vergleich von Graphen unter Berücksichtigung von beliebigen Annotationen an Knoten und Kanten ermöglicht. Die Evaluation verschiedener Annotationsmöglichkeiten für spezielle Anwendungen steht noch aus.

WEITERE SCHRITTE

Ein Schwerpunkt der zukünftigen Untersuchungen soll auf theoretischen Aspekten des MCS-Problems unter Berücksichtigung in der Praxis relevanter Nebenbedingungen liegen. Außerdem möchte ich das Thema *Graph Kernel* vertiefen, wozu externe Expertise im Bereich des maschinellen Lernens, speziell SVMs, hilfreich sein könnte. Die Verwendung von Ähnlichkeitsmaßen zwischen Molekülgraphen für die Clusteranalyse ist eine weitere interessante Anwendung. Hier könnte externe Expertise zu Clustering-Verfahren nützlich sein. Bzgl. der Anwendungen in der Chemie besteht eine Kooperation mit dem MPI für molekulare Physiologie in

Dortmund. Weitere externe Expertise zu möglichen Anwendungen der entwickelten Methoden ist wünschenswert.

REFERENCES

1. K. Klein, N. Kriege, and P. Mutzel. CT-index: Fingerprint-based graph indexing combining cycles and trees. In *IEEE 27th International Conference on Data Engineering (ICDE)*, pages 1115 –1126, april 2011.
2. J. W. Raymond, E. J. Gardiner, and P. Willett. Rascal: Calculation of graph similarity using maximum common edge subgraphs. *Comput. J.*, 45(6):631–644, 2002.
3. L. Schietgat, J. Ramon, M. Bruynooghe, and H. Blockeel. An efficiently computable graph-based metric for the classification of small molecules. In J.-F. Boulicaut, M. Berthold, and T. Horváth, editors, *Discovery Science*, volume 5255 of *Lecture Notes in Computer Science*, pages 197–209. Springer Berlin / Heidelberg, 2008.
4. S. V. N. Vishwanathan, N. N. Schraudolph, R. I. Kondor, and K. M. Borgwardt. Graph kernels. *Journal of Machine Learning Research*, 11:1201–1242, 2010.

Embedded Architecture Models

Marco Müller
paluno - The Ruhr Institute for Software Technology, Universität Duisburg-Essen
marco.mueller@uni-due.de
Prof. Dr. Michael Goedicke

PROBLEMBESCHREIBUNG UND FORSCHUNGSFRAGE

In komponentenbasierter Softwareentwicklung (CBSE) kann Softwarearchitektur mit Hilfe von Architekturbeschreibungssprachen modelliert werden. Zusätzlich wird die Architektur in Code und Konfiguration implementiert. Die Repräsentationen als Modell und Implementierung teilen sich verschiedene Sichten auf die Software. Dabei abstrahiert das Modell von der Implementierung, indem es Elemente auslässt, die als nicht Architekturrelevant gelten. Durch diese Aufteilung existieren zwei redundante Repräsentationen der Softwarearchitektur. Wenn diese beiden Repräsentationen nicht synchronisiert werden, können Missverständnisse entstehen, die zu erhöhtem Aufwand bei der Entwicklung und Wartung führen. Damit das Modell nicht langfristig von der Implementierung abweicht, muss die Synchronisierung in beide Richtungen möglich sein.

Diese Arbeit behandelt die Frage, ob die Redundanz der Repräsentationen für Softwarearchitekturen verhindert werden kann, indem die Modellierung mit der Implementierung integriert wird. Dazu sollen Architekturinformationen im Quellcode als statische Strukturen und Meta-Informationen eingefügt werden. Diese Informationen definieren zur Entwurfs- und Laufzeit die Architektur. Dadurch sind die *eingebetteten Architekturmodelle* die zentrale Repräsentation der Architektur, die für die Modellrepräsentation und die Implementierung gültig ist sind.

VORGEHENSWEISE UND METHODE

Ein geeignetes eingebettetes Architekturmodell und dazugehörige Tools werden experimentell entwickelt. Auf der Basis dieser Ergebnisse werden allgemeine Definitionen für die Einbettung von Architekturmodellen in Quellcode entwickelt. Die Ergebnisse werden in Praxisprojekten mit Studierenden evaluiert. Eine alternative Vorgehensweise wäre die Entwicklung eines allgemeinen Meta-Modells für eingebettete Architekturen gewesen, von dem einzelne Architekturmodelle abgeleitet werden könnten. Aufgrund des Fehlens eines Konsenses zur Definition von Softwarearchitektur wurde der experimentelle Ansatz vorgezogen.

VERWANDTE ARBEITEN

Die Arbeit ist aus den Zielen des CBSE [1] heraus motiviert, die Architektur explizit zu modellieren um die Komplexität von Software zu beherrschen. In dieser Arbeit soll die Architekturinformation nicht mehr implizit im Code sondern explizit vorliegen.

Die Entwicklung von lauffähigen Architekturen aus Architektursprachen ist mit Model Driven Architecture (MDA) [3] verwandt. Im Gegensatz zu MDA hat dieser Ansatz das Ziel, Redundanz auf der Architekturebene zu vermeiden. Bei MDA liegt die Architektur in den Modellen und dem Code redundant vor. Zudem fließen Änderungen an der generierten Architektur nicht in das Modell zurück.

Balz et al. [2] beschreiben die Einbettung von Modellen in Code. In ihren Arbeiten werden jedoch ausschließlich Verhaltensmodelle untersucht. Architekturbeschreibungen beschränken sich nicht auf Verhaltensmodelle, sondern betrachten auch z.B. Struktur- und Qualitätsmodelle.

VORLÄUFIGE ERGEBNISSE

Bisher wurden Verhaltensmodelle durch Schnittstellenautomaten in die Schnittstellenbeschreibung von Java integriert. Zur Laufzeit stehen Werkzeuge zum Prüfen und Durchsetzen der Verhaltensvorgaben im Komponentenmodell der Java Enterprise Edition zur Verfügung. Dadurch sind erste Ideen zum Einbetten von Kommunikations-Modellen umgesetzt.

NÄCHSTE SCHRITTE

Die Experimente sind bisher technisch erfolgreich. Der Nutzen muss noch durch Benutzerstudien untersucht werden. Die Sichten zur Architekturentwicklung mit dem Ansatz müssen noch ausgewählt, entwickelt und Konzepte zur Einbettung gefunden werden. Die konkreten Ergebnisse sollen soweit abstrahiert werden, dass Architekturen in beliebigen Architektursprachen eingebettet werden können. Dazu soll der Ansatz auch formalisiert werden. Zur Auswahl geeigneter Sichten und für Praxisprojekte sind zusätzliche Expertise oder Kooperationen wünschenswert.

REFERENCES

1. Szyperski, C.: Component Software: Beyond Object-Oriented Programming. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA (2002)

2. Balz, M., Striewe, M., Goedicke, M.: Continuous Maintenance of Multiple Abstraction Levels in Program Code. In: Proceedings of the 2nd International Workshop on Future Trends of Model-Driven Development - FTMDD 2010, Funchal, Portugal. (2010) 68–79

3. Brown, A.W., Iyengar, S., Johnston, S.: A Rational approach to model-driven development. IBM Systems Journal **45**(3) (2006) 463–480

Entwicklung neurobiologisch inspirierter Modelle für die Verarbeitung mehrkanaliger akustischer Signale

Dipl.-Inf. Axel Plinge
TU Dortmund
axel.plinge@tu-dortmund.de
Prof. Dr.-Ing. Gernot A. Fink

PROBLEMBESCHREIBUNG UND FORSCHUNGSFRAGE

In diesem Projekt werden erstmals komplexe Modelle der menschlichen akustischen Wahrnehmung für die Verarbeitung von mehrkanaligen Signalen von Mikrofonfeldern entwickelt. So wird für Objekte der auditiven Szene, wie etwas Sprecher in einem Konferenzraum, die Berechnung von *streams* zur Verfolgung im Raum und Sprachverbesserung möglich. Die innovative Methodik schlägt eine interdisziplinäre Brücke zwischen Psychophysik und Wahrnehmungspsychologie auf der einen Seite sowie Signalverarbeitung, Mustererkennung und effizienter Algorithmik auf der Anderen.

VORGEHENSWEISE UND METHODE

Hier werden sinnesphysiologisch, neurologisch und kognitiv plausible Modelle für mehrkanalige akustische Signale stufenweise von der Merkmalsextraktion auf mehrkanaligen Audiosignalen über die Segmentierung und Gruppierung bis zur modellbasierten sequentiellen Integration aufgebaut. Parallel werden schrittweise Demonstratoren erstellt. Die Lokalisierung mit immer komplexeren Modellen demonstriert jeweils den Fortschritt der Modellbildung. Ein Demonstrator zur Sprachtrennung wird nach der Implementierung der Gruppierung und auch in der Folge nach der Integration höherstufiger Modelle entstehen. Die Verfolgung unterschiedener oder identifizierter Sprecher mit aktiven Kameras ist ein mögliches abschliessendes Anwendugszenario.

VERWANDTE ARBEITEN

Als Modell für die Verarbeitung des menschlichen Gehirns beim Hörprozess hat sich die Theorie der auditiven Szenenanalyse [1], welche eine Vielzahl aus psychoakustischen Versuchen bekannte Phänomenen in einer weitgehend geschlossenen Theorie mit Anlehnung an die Gestalttheorie zusammenfasst, bewährt. Auf dieser Basis wurden verschiedenste (CASA) Computermodelle entwickelt [10]: Eine Realisierung lokalisationsbasierter Sprechertrennung verwendet eine Schätzung der Stimmtonhöhe um einkanalige Gruppierung zu realisieren und nutzt dann für die Itegration über die Zeit räumliche Merkmale [11]. Zur Sprecherunterscheidung werden aktuell die aus Langzeitumhüllenden in Gammaton-Filtern gewonnenen GFCCs als Merkmal verwendet [9]. In realen Umgebungen mit Störungen und signifikantem Hall existieren bislang kaum Anwendungen der (C)ASA. Diese Arbeit verwendet von Mikrofonfeldern statt ein oder zwei künstlichen Ohren, um in diesen Fällen bessere Ergebnisse zu erzielen. Es existiert eine Reihe technischer Verfahren für die bei der Lokalisierung und Signalverbesserung mit Mikrofonfeldern. Üblicherweise werden zur Sprachverbesserung bei mehrkanaligen Signalen von Sensorfeldern *beamforming*-Methoden eingesetzt, welche aus den vielen Kanälen ein einzelnes Signal mit einer räumlichen Richtcharakteristik bilden [5]. Ein populärer technischer Ansatz zur Sprechertrennung ist die *blind source separation* (BSS). Hier werden durch eine Kurzzeit-FFT berechnete spektrotemporale Teile des Signals über Verfahren wie die *independent componenent analysis* (ICA) Quellen zugeordnet, und die entsprechenden Teile zu den Signalen der einzelnen Quellen resynthetisiert [6].

VORLÄUFIGE ERGEBNISSE

In Dortmund wurde erstmals ein neurobiologisch und kognitionspyschologisch inspiriertes Modell entwickelt, welches die mehrkanaligen Signale eines Mikrofonfelds verarbeitet [8]. Mit einem innovativen, hallrobusten Modell der Cochlea, einem Verfahren zur integrierten Laufzeitschätzungen der Mikrofonpaare sowie einem auf Sprache ausgelegten Verfahren der höherstufigen Integration werden Sprecher in einem Konferenzraum lokalisiert. Es erwies sich robuster als das populäre GCC-PHAT Verfahren und auch als die Auswertung von Nulldurchgängen. Mit dieser Kombination von der Signalverarbeitung eines Mikrofonfelds mit einem einfachen neurobiologischen Modell konnte der Mehrwert einer solchen Lösung gezeigt werden. Das Modell wurde in [7] exemplarisch mit einem aktuellen Tracking-Algorithmus kombiniert. In der Folge wurde mit einer einfachen TTL-Heuristik [4] eine Echtzeitimplementierung in C++ realisiert. Die Lokalisierung und Verfolgung mehrerer gleichzeitig aktiver Sprecher in Echtzeit konnte in einem Konferenzraum mit starkem Hall ($T_{60} \approx 0.6\,\mathrm{s}$) demonstriert werden. Abbildung 1 zeigt ein Beipiel. Die fehlende Integration über Sprachpausen hinweg soll in der weiteren Entwicklung durch höherstufige Modelle erfolgen.

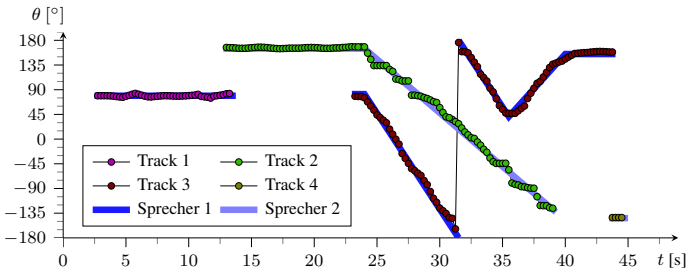

Figure 1. Tracking zweier Sprecher in einem Koferenzraum

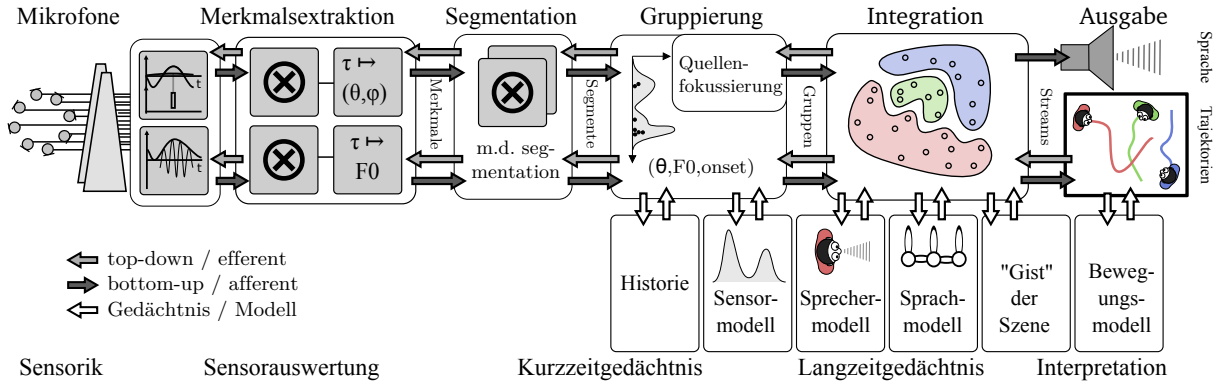

Figure 2. Architektur für ein sinnesphysiologisch, neurobiologisch und psychologisch plausibles Verarbeitungsmodell für mehrkanalige Signale

WEITERE SCHRITTE

Die zu entwickelnden Modelle werden sukzessive in *bottom-up*-Richtung aufgebaut, vgl. Abbildung 2. Dabei wird auch der *top-down*-Einfluss und die gegenseitige Beeinflussung der Verarbeitungsebenen genutzt. Der Einsatz von Methoden des maschinellen Lernens ermöglicht die Modellierung komplexer kognitiver Prozesse wie Adaptation, kontextbasierter Verarbeitung bis hin zur abstrakten Charakterisierung der auditiven Szene.

i) Zunächst werden Merkmalen aus den Signalen von einem oder mehreren Mikrofonfeldern extrahiert. Dabei ist insbesondere ein Verfahren für die Verwendung und Schätzung der Stimmtonhöhe auszuwählen und mit der Extraktion der anderen Merkmale in geschickter Weise zu kombinieren.

ii) Aus den Merkmalen werden Segmente im Zeit×Frequenz×Positions-Raum gebildet, welche anhand der Raumposition und Stimmtonhöhe gruppiert werden. Dabei ist ein effizientes Verfahren unter Verwendung von Mehrkernarchitekturen zu entwickeln.

iii) Nach der Gruppierung werden monaurale *simulateneous streams* für die Objekte der auditiven Szene berechnet. Dazu ist die Integration von *beamforming* [5] und *BSS* [6] Verfahren nötig, was auch in Kooperation geschehen kann. Für die übergreifende Kombination ist höchst wahrscheinlich das Clustering von Mischverteilungen mit dem bewährten EM-Algorithmus geeignet [2].

iv) Die Gruppen werden mit Hilfe von Sprecher- und Bewegungsmodellen über die Zeit zu *streams* einzelner Sprecher integriert. Für das Sprechermodell ist die Verwendung von GFCCs [9] und HMMs [3] vorgesehen. Das Bewegungsmodell kann mit CPHD-Filtern in weiterer Kooperation entwickelt werden [7].

v) Durch Integration eines Situations- und eines Aufmerksamkeitsmodells entstehen adaptive und reaktive Gesamtmodelle der akustischen Wahrnehmung.

REFERENCES

1. A. S. Bregman. *Auditory Scene Analysis*. MIT Press, 1990.
2. H. T. Dang Vu and R. Haeb-Umbach. An EM Approach to Integrated Multichannel Speech Separation and Noise Suppression. In *12th International Workshop on Acoustic Echo and Noise Control*, Tel Aviv, Israel, 2010.
3. G. A. Fink. *Markov Models for Pattern Recognition, From Theory to Applications*. Springer, Heidelberg, 2008.
4. N. Madhu and R. Martin. A Scalable Framework for Multiple Speaker Localization and Tracking. In *11th International Workshop on Acoustic Echo and Noise Control*, Seattle, Washington USA, 2008.
5. R. Martin, U. Heute, and C. Antweiler, editors. *Advances in Digital Speech Transmission*. Wiley, 1 edition, 2008.
6. F. Nesta, P. Svaizer, and M. Omologo. Convolutive BSS of Short Mixtures by ICA Recursively Regularized Across Frequencies. *IEEE Transactions on Audio, Speech and Language Processing*, 19(3):624–639, Mar. 2011.
7. A. Plinge, D. Hauschildt, M. H. Hennecke, and G. A. Fink. Multiple speaker tracking using a microphone array by combining auditory processing and a gaussian mixture cardinalized probability hypothesis density filter. In *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, Prague, Czech Republic, 2011.
8. A. Plinge, M. H. Hennecke, and G. A. Fink. Robust neuro-fuzzy speaker localization using a circular microphone array. In *Proc. 12th International Workshop on Acoustic Echo and Noise Control*, Tel Aviv, Israel, 2010.
9. S. O. Sadjadi and J. H. L. Hansen. Hilbert Envelope based Features for robust Speaker Idetification under reverberant mismatched Conditions. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 5448–5451, 2011.
10. D. Wang and G. J. Brown, editors. *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*. IEEE Press / Wiley, 2006.
11. J. Woodruff and D. Wang. Sequential Organization of Speech in Reverberant Environments by Integrating Monaural Grouping and Binaural Localization. *IEEE Transactions on Audio, Speech and Language Processing*, 18(7):1856–1866, Sept. 2010.

Videobasierte Gestenerkennung in einer intelligenten Umgebung

Jan Richarz
Fakultät für Informatik, LS XII, TU Dortmund
jan.richarz@udo.edu
Betreuer: Prof. Dr.-Ing. Gernot A. Fink

PROBLEMBESCHREIBUNG UND FORSCHUNGSFRAGE

Die Arbeit umfasst die Konzeption und prototypische Umsetzung einer berührungslosen und nutzerunabhängigen visuellen Klassifikation von Armgesten anhand ihrer räumlich-zeitlichen Bewegungsmuster mit Methoden der *Computer Vision* und der Mustererkennung. Das Anwendungsszenario ist die Steuerung von Geräten in einem intelligenten Raum. Die Besonderheit der Arbeit besteht darin, dass keine starken Einschränkungen hinsichtlich des Interaktionsszenarios getroffen werden und die Analyse somit in einem Multikamera-Szenario mit prinzipiell beliebiger Kamerakonfiguration vorgenommen wird. Insbesondere sind die Kameras unsynchronsiert und ggf. aktiv, was spezielle Anforderungen an die Adaptivität und Robustheit der verwendeten Methoden zur Folge hat.

VORGEHENSWEISE UND METHODE

Meine Dissertation ist eine Ingenieursarbeit, die auf etablierten Methoden aufsetzt, diese in einem realen Anwendungsumfeld erprobt, gezielt verbessert und zu einem robusten wie reaktiven Gesamtsystem kombiniert. Demzufolge ist der theoretische Anteil gering, der experimentelle und empirische Anteil groß. Ein besonderer Fokus liegt auf der Nähe zur konkreten Anwendung, weshalb die prototypische Umsetzung aller erarbeiteten Ansätze einen großen Stellenwert einnimmt. Zentrale Methoden der Arbeit sind statistische Modellierung der Farb- und Kantenverteilungen der Kamerabilder, Klassifikation von Bildbereichen mit künstlichen neuronalen Netzen sowie die Analyse räumlich-zeitlicher Trajektorien von Körperteilen mit Hidden Markov Modellen (HMM). Aus der Fülle möglicher Methoden wurden diese aufgrund ihrer bekannten Leistungsfähigkeit und effizienten algorithmischen Umsetzbarkeit ausgewählt.

VERWANDTE ARBEITEN

Im Bereich der Gesten- und Aktionserkennung existiert eine Vielzahl interessanter Forschungsarbeiten. Thematisch sehr nah verwandt sind z.B. die Arbeiten von Alon et al. [1] und Wang et al. [2]. In [1] wird für die Detektion von Körperteilen ein meiner Arbeit sehr ähnlicher Ansatz verfolgt und ein interessantes Konzept für das automatische Lernen von Subgesten-Relationen vorgestellt. Die Autoren beschränken sich jedoch auf ein stark eingeschränktes, monokulares Szenario. Im Gegensatz dazu wird in [2] eine Stereokamera verwendet und die zusätzliche Tiefeninformation zum *Tracking* eines Körpermodelles genutzt. Die Gelenkwinkel dieses Modelles bilden die Grundlage für die Gestenanalyse. Auch dieser Ansatz ist jedoch nicht ansichtsinvariant, weil die Person sich annähernd frontal vor der Stereokamera befinden muss. In dieser Hinsicht stellen die Verwendung eines echten Multikamerasystems mit beliebiger Kameraanordnung – besonders die Verwendung unsynchronisierter Kameras – sowie die Annahme eines weitgehend uneingeschränkten, ansichtsinvarianten Szenarios die wichtigsten Neuerungen meiner Arbeit dar. Als weitere sehr interessante Arbeit ist [3] zu nennen. Die Autoren verfolgen zwar ein anderes Ziel – die Realisierung ubiquitärer Touchscreens an beliebigen Flächen im Raum mittels eines Systems aus mehreren Tiefenkameras – zeigen aber eindrucksvoll die Möglichkeiten, die sich durch alternative Interaktionskonzepte ergeben.

ERGEBNISSE

Es wurde eine komplette parallele Verarbeitungspipeline für die einzelnen Kamerabildströme entwickelt, in der Personen detektiert, *getrackt* und die Positionen ihrer Hände robust gefunden werden. Die Evaluation ergab hierfür sehr gute Ergebnisse unter realistischen und veränderlichen Innenraumbedingungen. Der Prototyp läuft beinahe in Echtzeit. Darauf aufbauend wurden Module für die 3D-Kombination der Ergebnisse mehrere Kameras unter Beachtung des zeitlichen Versatzes bei der Aufnahme, die Bewertung und Aggregation der 3D-Hypothesen zu Trajektorien sowie die Erkennung dieser mit HMM realisiert. Die Evaluation des Gestenerkenners auf einer Datenbank mit neun verschiedenen Armgesten zeigt vielversprechende Ergebnisse.

WEITERE SCHRITTE

Die Dissertation wurde im September 2011 eingereicht.

REFERENCES

1. Alon, J.; Athitsos, V.; Yuan, Q.; Sclaroff, S.: A unified framework for gesture recognition and spatiotemporal gesture segmentation. In: *IEEE Trans. PAMI* 31 (2009), Nr. 9, S. 1685–1699.

2. Wang, S. B.; Quattoni, A.; Morency, L.-P.; Demirdjian, D.: Hidden conditional random fields for gesture recognition. In: *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*. 2006.

3. Wilson, A. D.; Benko, H.: Combining multiple depth cameras and projectors for interactions on, above and between surfaces. In: *Proc. ACM Symp. on User Interface Software and Technology*. 2010.

Learning bag-of-features representations for handwriting recognition

Leonard Rothacker
TU Dortmund, Fakultät für Informatik
leonard.rothacker@tu-dortmund.de
Prof. Dr.-Ing. Gernot A. Fink

PROBLEMBESCHREIBUNG UND FORSCHUNGSFRAGE

Verfahren zur automatischen Handschrifterkennung gliedern sich grob in die Schritte Vorverarbeitung, Merkmalsextraktion und Erkennung. Dabei werden bei der Vorverarbeitung und Merkmalsextraktion viele Heuristiken verwendet, die sich experimentell als günstig erwiesen haben. Rechtfertigen lassen sich die Entscheidungen für die entsprechenden Verfahren und deren Parameter somit nur aufgrund von Expertenwissen. Ein Problem dabei ist, dass sich Ansätze so nicht mehr ohne weiteres über verschiedene Problemfelder hinweg einsetzen lassen. Insbesondere lassen sich jedoch mit heuristischen Verfahren keine Merkmale extrahieren, die systematisch von der Variabilität der Daten abstrahieren und so die Erkennungsleistung verbessern.

Ziel der Forschung ist entsprechend das automatische Erlernen einer Merkmalsrepräsentation, wobei der Lernprozess durch die Erkennungserbenisse gesteuert sein soll. Es soll gezeigt werden, dass große Variabilitäten weniger starke Auswirkungen haben und sich Merkmalsrepräsentationen für Handschriften unterschiedlicher Sprachen automatisch erlernen lassen.

VORGEHENSWEISE UND METHODE

Zur automatischen Handschrifterkennung werden Hidden Markov Modelle (HMMs) erfolgreich eingesetzt [3]. Die Erkennungshypothesen, die sich auch während des Trainings durch das HMM ableiten lassen, sollen nun zum Erlernen einer Merkmalsrepräsentation verwendet werden. Da die Hypothesen auch maßgeblich von den Merkmalen abhängen, ergibt sich eine wechselseitige Beziehung, die in einen iterativen Prozess der Aktualisierung von Merkmalen und Generierung von Hypothesen integriert werden soll.

Für das Erlernen von Merkmalsrepräsentationen sollen lokale Bilddeskriptoren verwendet werden. Dabei werden die Deskriptoren in einen Unterraum projiziert, in dem Deskriptoren einer Klasse nah beieinander und Deskriptoren verschiedener Klassen weit auseinander liegen. Die Klassenzugehörigkeiten sollen aus den HMM Hypothesen abgeleitet werden. Zur Projektion in einen Unterraum bieten sich lineare (Linear Discriminant Embedding) und nichtlineare (Autoencoder) Verfahren an [2].

Zur Integration von HMMs mit lokalen Bilddeskriptoren wird ein Bag-of-Features Ansatz [1] verwendet. Deskriptoren werden dabei zu einem visuellen Vokabular (den Features) quantisiert. Eine Statistik über die vorkommenden Features dient als Beschreibung eines Bildes. Für das HMM wird die Beschreibung je Frame erzeugt. Die Wahrscheinlichkeiten einzelner Features bezüglicher der Zustände in dem HMM lassen sich trainieren.

Zur Verifikation der Ergebnisse sollen die IAM Datenbank (lateinische Schrift) und die IFN/ENIT Datenbank (arabische Schrift) verwendet werden.

VERWANDTE ARBEITEN

In [1] werden Bag-of-Features Repräsentationen zur Bildsuche (image retrieval) eingesetzt. [2] erweitert den Ansatz um ein Verfahren zum automatischen Lernen von Deskriptoren. Für Handschrifterkennung [3] werden geometrische und einfache analytische Merkmale mit HMMs verwendet. Die Neuerung besteht nun in der Verwendung von HMMs mit Bag-of-Features Repräsentationen, wobei die Deskriptoren gelernt werden.

VORLÄUFIGE ERGEBNISSE

Die Integration des HMM mit dem Bag-of-Features Ansatz ist erfolgt und es konnte gezeigt werden, dass gleiche Ergebnisse wie mit geometrischen Merkmalen möglich sind. Es wurden dazu die IAM-Datenbank sowie die IFN/ENIT Datenbank verwendet.

WEITERE SCHRITTE

Als nächstes muss ein Ansatz zum Erlernen von Deskriptoren integriert werden. Die wichtigsten Fragen dabei sind:

- Wie lassen sich die Erkennungshypothesen zur Annotation der Deskriptoren verwenden?
- Wie wirkt sich das geänderte visuelle Vokabular in der nächsten Iteration im HMM Training aus?

REFERENCES

1. S. O'Hara and B. A. Draper. Introduction to the bag of features paradigm for image classification and retrieval. *Computing Research Repository*, arXiv:1101.3354v1, 2011.
2. J. Philbin, M. Isard, J. Sivic, and A. Zisserman. Descriptor learning for efficient retrieval. In *Proceedings of the European Conference on Computer Vision*, 2010.
3. T. Plötz and G. A. Fink. Markov Models for Offline Handwriting Recognition: A Survey. *International Journal on Document Analysis and Recognition*, 12(4):269–298, 2009.

Eingabesensitive Approximation von Flächen und deren Eigenschaften, basierend auf Punktwolken

Christian Scheffer
Fachbereich Informatik und Technische Universität Dortmund
christian.scheffer@tu-dortmund.de
Betreuer: Prof. Dr. Jan Vahrenhold

PROBLEMBESCHREIBUNG UND FORSCHUNGSFRAGE

Es wird allgemein die Fragestellung behandelt, wie und wie gut (Ober-)Flächen im $\mathbb{R}^3$ und ihre Eigenschaften lediglich von einer Punktwolke ausgehend rekonstruiert bzw. approximiert werden können. Es sollte einleuchtend sein, dass mit steigendem Anspruch an die Qualität der Berechnungen die Dichte der Punktmenge steigen muss. Diese Problemstellung findet viele Anwendungen in der Praxis. Hier wäre zum Beispiel die dreidimensionale Mustererkennung zu nennen. Das Vergleichen der approximierten Eigenschaften zweier zu Grunde liegenden Flächen ausgehend von zwei entsprechenden Punktwolken könnte zu einer eingabesensitiven Vergleichsmethode führen, deren Entscheidungsfehler mit steigender Eingabequalität sinkt. Mit der Entwicklung immer genauerer und leistungsfähigerer Geräte, die die betrachteten Eingabepunktmengen generieren, wird die oben beschriebene Problemstellung umso praxisrelevanter.

VORGEHENSWEISE UND METHODE

Wie bereits oben erwähnt, wird als Eingabe eine Punktmenge vorausgesetzt, die auf der betrachteten bzw. zu approximierenden Fläche liegt. Selbst für beliebig dichte Punktwolken kann gezeigt werden, dass sie sogenannte *gültige Samplemengen* für unterschiedliche Flächen sind. Diese Unterschiede sind jedoch mit steigender Sampledichte verschwindend gering, sodass es das Ziel sein muss, Approximationsalgorithmen zu entwickeln, deren Ausgabefehler antiproportional von der Eingabequalität abhängen. Diese gewünschte Eigenschaft soll durch multiplikative Approximationsfehler bewiesen werden. Als Grundlage für entsprechende Beweise muss zunächst ein mathematisches Kalkül definiert werden, das ein *gültiges Sample* beschreibt, ein sogenanntes ε-Sample. Solch eine Punktdiskretisierung berücksichtigt durch erhöhte Punktdichte feinere Details, d.h. stärkere Krümmung oder Faltung der zu Grunde liegenden Fläche. Ausgehend von diesen mathematisch beschreibbaren Voraussetzungen lassen sich dann quantitative Analysen der entworfenen Methoden durchführen. Zusätzlich werden visuelle Inspektionen der Ausgaben durchgeführt, die als Unterstützung und Veranschaulichung der Forschungsergebnisse dienen.

VERWANDTE ARBEITEN

Eines der prominentesten Ergebnisse auf dem oben beschriebenen Arbeitsgebiet ist der sogenannte *CoCone* Algorithmus, entworfen von Amenta *et al.* [3]. Dieser rekonstruiert beweisbar korrekt unberandete glatte Flächen im $\mathbb{R}^3$. Die Korrektheit wird bewiesen, indem Amenta *et al.* zeigen, dass ihre Flächenrekonstruktion topologisch äquivalent zur Ursprungsoberfläche ist. Somit sollte der allgemeine Anspruch weiterer Rekonstruktionsmethoden sein, zu beweisen, dass die Topologie der Ausgangsfläche nicht verfälscht wird. Eine weitere wichtige Arbeit ist ein Oberflächenrekonstruktionsalgorithmus [1], der einen fast linearen Zeitverbrauch hat. Speziell wird eine Submethode vorgestellt, mit der die Ausgangspunktmenge so ausgedünnt wird, dass das resultierende Sample immer noch dicht genug ist, aber eine lokale Gleichmässigkeit bzgl. Dichte aufweist. Diese Ausdünnungsmethode ist einer der Grundbausteine der Arbeiten dieses Forschungsvorhabens. Als Letztes sei noch ein Algorithmus [2] genannt, der ursprünglich aus einem anderem Teilgebiet der Algorithmischen Geometrie kommt. Dieser berechnet auf einer stückweise linearen geschlossenen Fläche kürzeste Wege in optimaler Zeit. Auch dieser Algorithmus wurde beim Entwurf neuer Methoden als *Black Box* benutzt.

VORLÄUFIGE ERGEBNISSE

Das erste Ergebnis dieser Forschungsarbeit ist ein Algorithmus, der angelehnt an [1] Flächen rekonstruieren kann, die einen glatten Rand haben. Hier gab es bisher lediglich eine Arbeit [4], die einen ähnlichen Algorithmus vorstellt, der jedoch mehr Informationen als der entworfene Algorithmus als Eingabe benötigt.

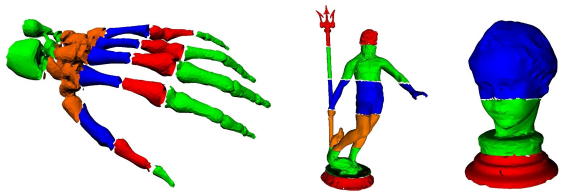

Figure 1. Rekonstruktionen von in Schichten geschnittene Flächen

Als zweites ist eine ebenfalls bereits entwickelte Methode zu nennen, die in fast optimaler Zeit eingabesensitiv die Länge der kürzesten Kurven, die sogenannten *Geodäten*, auf der betrachteten Fläche approximieren kann. Beide Eigenschaften wurden mathematisch nachgewiesen, allerdings noch nicht experimentell getestet, sodass eine experimentelle Evaluation ein nächster Schritt sein könnte.

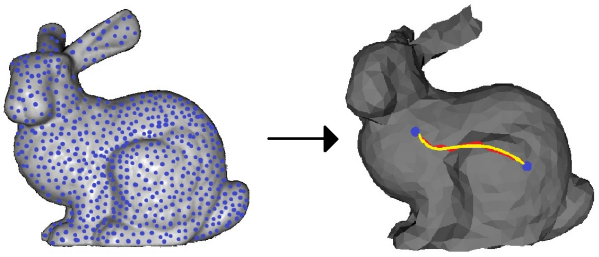

Figure 2. Approximation von Kurvenlängen auf der zu Grunde liegenden Fläche bzw. dessen Rekonstruktion, ausgehend von einer Punktmenge auf dieser

WEITERE SCHRITTE

Mittelfristig sollen weitere Algorithmen entworfen werden, die zusätzliche Eigenschaften der betrachteten Fläche approximieren, sodass ein Repertoire von Methoden entsteht, die es ermöglichen im Sinne der Mustererkennung zwei Flächen miteinander zu vergleichen und basierend auf den approximierten Eigenschaften eine "Ähnlichkeitsentscheidung" zu treffen.

REFERENCES

1. S. Funke and E. A. Ramos. Smooth-surface reconstruction in near-linear time. In *Proc. Symposium on Discrete Algorithms*, pages 781–790. ACM, 2002.
2. Y. Schreiber. An optimal-time algorithm for shortest paths on realistic polyhedra. *Discrete & Computational Geometry*, 43:21–53, 2010.
3. N. Amenta and M. Bern. and S. Choi. and T. K. Dey. and N. Leekha. A simple Algorithm for Homeomorphic Surface Reconstruction. In *Proc. Symposium on Computational Geometry*, pages 213–222. ACM, 2000.
4. T. K. Dey and K. Li and E. A. Ramos and R. Wenger Isotopic Reconstruction of Surfaces with Boundaries. *Computer Graphics Forum*, 28(5):1371–1382, 2009.

Spezielle Clusteringprobleme

Melanie Schmidt
Fakultät für Informatik, TU Dortmund
melanie.schmidt@udo.edu
Prof. Dr. Christian Sohler

PROBLEMBESCHREIBUNG

Grob gesprochen unterteilt ein *Clustering* eine Datenmenge in Gruppen, wobei Objekte in aus der gleichen Gruppen ähnlich sind und Objekte aus verschiedenen Gruppen sich voneinander deutlich unterscheiden. In meinem Promotionsvorhaben geht es um verschiedene Variationen von Clustering anhand von (quadrierter) euklidischer Distanz bei vorgegebener Anzahl von Clustern (Gruppen). Für eine gegebene Punktemenge im Euklidischen Raum besteht eine Lösung für dieses Problem aus der Angabe von k Clusterzentren, und die Kosten einer Lösung ist die Summe der quadrierten Abstände zum jeweils nächsten Clusterzentrum. Die Minimierung dieser Kostenfunktion ist als k-means Problem bekannt und in seiner grundlegenden Fragestellung sehr gut untersucht. In der Praxis tritt Clustering aber nicht zwangsläufig in der in der Theorie üblicherweise untersuchten Standardform auf. Daher interessieren wir uns für Erweiterungen.

VORGEHENSWEISE UND METHODE

Mein Promotionsvorhaben ist im Bereich der theoretischen Informatik angesiedelt und somit hauptsächlich theoretischer Natur. Aufgrund von Verbindungen zum Sonderforschungsbereich 876 und dem Schwerpunktprogramm „Algorithm Engineering“ sind aber auch experimentelle Anteile möglich und angedacht.

VERWANDTE ARBEITEN UND FORSCHUNGSFRAGEN

Zwei der von uns betrachteten Erweiterungen sind durch verwandte Arbeiten motiviert. Der k-means Algorithmus ist der in der Praxis am häufigsten verwendete Algorithmus für das oben beschriebene k-means Problem. Er wurde von Lloyd [3] bereits 1982 entwickelt. Vergleichsweise kürzlich entwickelten Arthur und Vassilvitskii [1] eine Initialisierungsmethode für diesen Algorithmus, die zum ersten Mal zu einer beweisbaren Gütegarantie für den k-means Algorithmus führt. Die Initalisierungsmethode basiert auf einem cleveren adaptiven Samplingalgorithmus. Wir würden ihre Ideen gerne auf etwas weiter gefasste Fragestellungen übertragen, zum Beispiel auf EM-artige Algorithmen.

Bei *probabilistischem Clustering* sind die Punkte in der Eingabepunktemenge nicht mehr sicher an einem bestimmten Punkt im euklidischen Raum, sondern können sich mit gewisser Wahrscheinlichkeit an unterschiedlichen Stellen befinden. Solche Daten erhält man z.B. durch Sensormessungen. Minimiert werden dann die erwarteten Kosten. Die bisher einzige theoretische Arbeit zu diesem Thema stammt von Cormode und McGregor [2], die für einige Varianten von Clusteringproblemen zeigen, wie man bekannte Algorithmen weiterhin nutzen kann. Nicht gelöst wurde das Euklidische k-median Problem, bei dem im Gegensatz zum k-means Problem einfache euklidische Abstände verwendet werden.

VORLÄUFIGE ERGEBNISSE

Das probabilistische k-median Problem ist besonders schwierig, wenn man zusammen mit der Berechnung der Clusterzentren bereits festlegen muss, welcher Eingabepunkt zu welchem Zentrum gehören soll, ohne dessen späteren Aufenthaltsort zu kennen. Wir haben für diesen Fall einen Algorithmus entwickelt, der auf sogenannten Kernmengen basiert. Eine Kernmenge ist eine gewichtete Teilmenge der Eingabedaten, die die Eingabe zusammenfasst, d.h. die für jede Wahl von k Clusterzentren eine gute Approximation für die Kosten liefert.

WEITERE SCHRITTE

Momentan beschäftige ich mich vor allem mit der zuerst genannten Forschungsfrage. Wir versuchen, die Strategie von Arthur und Vassilvitski auf EM-Algorithmen für sphärische Gaußverteilungen zu übertragen. Hier haben wir bereits etliche Erkenntnise gewonnen, konnten aber noch keinen beweisbar guten Algorithmus für das Problem entwickeln. Wir haben bereits eine einfache Implementierung unserer Algorithmenideen durchgeführt, die wir gerne weiter verfeinern möchten. Durch verschiedene Experimente können wir möglicherweise Rückschlüsse darüber ziehen, welche Algorithmenideen wir theoretisch näher untersuchen sollten.

REFERENCES

1. Arthur, David und Sergei Vassilvitskii. *k-means++: The Advantages of Careful Seeding*, Proceedings of SODA 2007, S. 1027–1035. 2007

2. Cormode, Graham und Andrew McGregor. *Approximation algorithms for clustering uncertain data*, Proceedings of PODS 2008, S. 191–200. 2008

3. Lloyd, Stuart P. *Least Squares Quantization in PCM*, IEEE Transactions on Information Theory 28, S. 129–137. 1982

Streamingalgorithmen für diskriminative Modelle

Chris Schwiegelshohn
Fakultät Informatik, TU Dortmund
chris.schwiegelshohn@udo.edu
Christian Sohler

PROBLEMBESCHREIBUNG

Klassifikation ist ein zentrales Problem im Bereich des Machine Learning. In meiner Dissertation beschäftige ich mich mit Approximationsalgorithmen für Klassifikatoren, zu denen die logistische Regression gehört. Für eine Klassifikation mit zwei Klassen, sind die *Odds* definiert als das Verhältnis $\frac{P(Y=1)}{P(Y=0)} = \frac{P(Y=1)}{1-P(Y=1)}$, bzw im diskriminativen Fall als $\frac{P(Y=1|\mathbf{X})}{1-P(Y=1|\mathbf{X})}$ für eine Menge an beobachteten Merkmalen $\mathbf{X}$. Ergeben die Odds einen Wert größer 1 so ist die 1-Klasse wahrscheinlicher, ergeben sie einen Wert kleiner 1 so ist 0-Klasse wahrscheinlicher. Die Modellannahme der logistischen Regression besteht daran, das der Logarithmus der Odds eine lineare Funktion von $\mathbf{X}$ ist, also $\log \frac{P(Y=1|\mathbf{X})}{1-P(Y=1|\mathbf{X})} = \mathbf{w}^{\mathrm{T}}\mathbf{X}$. Eine geeignete Wahl der Parameter sorgt für ein Modell, welches die vorliegenden, bereits klassifizierten Rohdaten möglichst gut erklärt. Generell wird der Parametervektor $\mathbf{w}$ für einen gegebenen Datensatz mit N Daten durch Maximum Likelihood bestimmt [5]. Fällt N jedoch sehr groß aus, müssen approximative Methoden herangezogen werden, die den Datensatz auf eine Größe $n << N$ verkleinern und dabei trotzdem einen Parametervektor $\mathbf{w}_n$ bestimmen, der in etwa dem theoretischen Optimum $\mathbf{w}_N$ entspricht.

VORGEHENSWEISE

Meine Arbeit hat zunächst einen großen theoretischen Anteil, die resultierenden Algorithmen können und sollen allerdings auch experimentell evaluiert werden. Referenzalgorithmen zu der untersuchten Fragestellung sind mir zur Zeit nicht bekannt.

VERWANDTE ARBEITEN

Das d-dimensionale lineare Regressionsproblem besteht darin, zu einer gegeben Menge an N Daten mit Beobachtungen $\mathbf{A} \in \mathbb{R}^{N\times d}$ und Wertevektor $\mathbf{b} \in \mathbb{R}^N$ Parameter $\mathbf{w} \in \mathbf{R}^d$ zu finden, so dass $||Aw - b||_2$ minimiert wird. Geometrisch entspricht w einer Hyperebene mit minimalem euklidischen Abstand zu allen Datenpunkten. Bei einer großen Anzahl an Daten lassen sich Approximationen durch das Einführen einer Auswahlmatrix $\mathbf{S} \in \mathbb{R}^{N\times k}$ erzielen, wobei k die neue, reduzierte Problemgröße darstellt. Es gibt einige Resulte, die Schranken für k finden, sodass die optimale Lösung von $||\mathbf{S}^{\mathrm{T}}(Aw - b)||_2$ in etwa der optimalen Lösung von $||Aw - b||_2$ entspricht. Die Rolle von $\mathbf{S}$ kann dabei sehr unterschiedlich sein. Drineas et.al [4] gaben ein Verfahren an, in dem $O(d^2)$ Zeilen von $\mathbf{A}$ und $\mathbf{b}$ mit nicht uniformen Wahrscheinlichkeiten ausgewählt wurden. Im Prinzip sind dadurch die Berechnung approximativer Lösungen möglich, die von der Größe des Datensatzes nicht mehr abhängen. Allerdings geht N in die Auswahlwahrscheinlichkeiten ein und es zur Zeit noch unbekannt ob sich diese schneller berechnen lassen als sich das Ursprungsproblem lösen lässt. Weiterere Ansätze [1][2] nutzen das Johnson-Lindenstrauss-Lemma [3] für eine Aggregation der Daten mittels Einbettung.

VORLÄUFIGE ERGEBNISSE

Eine direkte Übertragung der Ergebnisse aus der linearen Regression erwies sich als sehr schwierig, da diese auf lineare Algebra basieren, welche sich nicht auf das logistische Regressionsmodell anwenden lässt. Eine Linearisierung hat an dieser Stelle erste Resultate erzielt.

WEITERE SCHRITTE

Kurzfristig sollen durch die Linearierung anfallenden Laufzeiteinbußen weiter gesenkt werden, ferner steht eine experimentelle Evaluierung der Algorithmen noch an. Langfristig werde ich die Übertragbarkeit der Resultate auf allgemeinere Modelle wie den Conditional Random Fields [6] untersuchen.

REFERENCES

1. T. Sarlós, Improved approximation algorithms for large matrices via random projections. In *Proceedings of the 47th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, 143-152, 2006.

2. K. L. Clarkeson and D. P. Woodruff, Numerical linear algebra in the streaming model. In *Proceedings of the 41th ACM Symposium on Theory of Computing (STOC)*, 341-350, 2010.

3. W. B. Johnson and J. Lindenstrauss, Extensions of Lipschitz mappings into a Hilbert Space. Contemporary Mathematics 26: 189-206, 1984.

4. P. Drineas, M. W. Mahoney, and S. Muthukrishnan, Sampling algorithms for ℓ_2 regression and applications. In *Proceedings of the 17th Symposium on Discrete Algorithms (SODA)*, 1127-1136, 2006.

5. C. M. Bishop, 0387310738, Springer, Pattern Recognition and Machine Learning (Information Science and Statistics), 2007.

6. C Sutton and A McCallum, An Introduction to Conditional Random Fields for Relational Learning. In *Introduction to Statistical Relational Learning*, edited by Lise Getoor and Ben Taskar, 2006.

Entwurf und Analyse ressourceneffizienter Lernverfahren

Sylvie Temme
Fakultät für Informatik, Technische Universität Dortmund
Sylvie.Temme@tu-dortmund.de
Betreuer: Prof. Dr. Jan Vahrenhold

PROBLEMBESCHREIBUNG UND FORSCHUNGSFRAGE

Die Arbeit findet im Rahmen des Teilprojekts A2 des Sonderforschungsbereichs 876 statt. Ziel des Teilprojekts ist die theoretische Untersuchung der Algorithmik von Lernverfahren in eingebetteten Systemen. Dazu sollen wesentliche Aspekte von eingebetteten Systemen wie Energiebedarf oder Bandbreite neben den klassischen Komplexitätsmaßen wie Laufzeit und Speicherbedarf bei der Bewertung von Algorithmen berücksichtigt werden. Dies erfordert zunächst die Entwicklung neuer Modelle zum Algorithmenentwurf. Der Entwurf und die Analyse von Algorithmen wird klassischerweise im *real*-RAM-Modell vorgenommen, in dem unter anderem angenommen wird, dass die Zugriffszeit auf jede Speicherzelle gleich ist. Oft werden Algorithmen aber unter Ressourcenbeschränkungen eingesetzt, Ressourcen (z.B. Speichergröße oder Energie) stehen nur in geringem Verhältnis zur Größe der zu verarbeitenden Eingabe zur Verfügung.

In Computerarchitekturen tritt beispielsweise eine Speicherbeschränkung an mehreren Stellen auf, da dort mehrstufige Speicherhierarchien eingesetzt werden. Die Daten können dort nur in einem kleinen, schnellen Speicher verarbeitet werden, während sie wegen ihrer Größe in einem großen, langsameren Speicher abgelegt werden müssen. Beim Umgang mit großen Datenmengen ist daher oft die Zeit, die benötigt wird, um Daten aus dem größeren in den kleineren Speicher nachzuladen, dominierend. Aggarwal und Vitter [2] haben das „Externspeichermodell“ entwickelt, das eine zweistufige Hierarchie modelliert, Frigo *et al.* [3] führten das *cache-oblivious*-Modell als ein allgemeineres Berechnungsmodell für mehrstufige Speicherhierarchien ein.

Im Vordergrund des Projekts steht die Entwicklung und Analyse von Algorithmen für eingebettete Systeme unter dem Aspekt des Energiebedarfs. Als Startpunkt für die Untersuchungen soll dort zunächst eine Erweiterung des Externspeichermodells dienen, das um den Begriff des Energiebedarfs erweitert wird. Ziel ist die Analyse und algorithmische Umsetzung klassischer Lernverfahren für dieses und weiterführende Modelle. Dabei soll neben Laufzeit, Speicher- und Energiebedarf auch ggf. die Approximationsgüte des Algorithmus ein wichtiges Kriterium der Bewertung von Algorithmen darstellen.

VORGEHENSWEISE UND METHODE

Das Dissertationsvorhaben ist als theorieorientiertes Vorhaben ausgelegt. Es ist jedoch geplant, einzelne Verfahren auch experimentell zu evaluieren.

VERWANDTE ARBEITEN

Frigo *et al.* [3] definieren das *cache-oblivious*-Modell und stellen grundlegende Algorithmen wie z.B. Sortieren in diesem Modell vor. Vitter [5] bietet eine Übersicht über den aktuellen Stand der Forschung im Bereich „Externspeicheralgorithmen“. Gieseke *et al.* [4] beschreiben ressourceneffiziente Bausteine, die wir benutzt haben.

VORLÄUFIGE ERGEBNISSE

Wir haben einen Ansatz von Abam und Har-Peled [1] zur Berechnung einer SSPD (*semi-separated pair decomposition*; eine Datenstruktur zur Datenanalyse) so modifiziert, dass er ressourceneffizient ist. Formal konnten wir nachweisen, dass unser Algorithmus im *cache-oblivious*-Modell nur $O\left(\frac{N}{B\varepsilon^d}\log_{M/B}\frac{N}{B\varepsilon^d}\log_2 N\right)$ Speichertransfers benötigt. Hierbei bezeichnet N die Eingabegröße, ε einen Güteparameter, d die Dimension des Datenraums, M die Größe des kleinen Speichers und die B Anzahl der Daten, die mit einem Speichertransfer in den kleinen Speicher geladen werden.

WEITERE SCHRITTE

Das Dissertationsvorhaben befindet sich in der Anfangsphase. Ausgehend von den obigen Ergebnissen sollen Algorithmen, die auf der Verwendung einer SSPD basieren [1], auf ihre *cache*-Effizienz hin untersucht werden.

REFERENCES

1. Abam, Mohammad A.; Har-Peled, Sariel: New constructions of SSPDs and their applications. In: *Proc. Symp. Computational Geometry* 2010, 192–200.
2. Aggarwal, Alok; Vitter, Jeffrey S.: The input/output complexity of sorting and related problems. *Comm. ACM* 31:1116–1127, 1988.
3. Frigo, Matteo; Leiserson, Charles E.; Prokop, Harald; Ramachandran, Sridhar: Cache-Oblivious Algorithms. In: *Proc. Symp. Foundations of Computer Science* 1999, S. 285–298.
4. Gieseke, Fabian; Gudmundsson, Joachim; Vahrenhold, Jan: Pruning spanners and constructing well-separated pair decompositions in the presence of memory hierarchies. *J. Discr. Algorithms* 8:259–272, 2010.
5. Vitter, J. S.: *Algorithms and Data Structures for External Memory*. Foundations and Trends in Theoretical Computer Science, 2(4):305-474, 2008.

Progressive Algorithmen und Datenstrukturen zur Datenanalyse unter Ressourcenbeschränkungen mit Anwendungen in der Astronomie

Andreas Thom
Fakultät für Informatik - TU Dortmund
andreas.thom@tu-dortmund.de
Betreuer: Prof. Dr. Jan Vahrenhold

PROBLEMBESCHREIBUNG UND FORSCHUNGSFRAGE

Die Survey-Astronomie ist ein besonders schnell wachsender Zweig der beobachtenden Astronomie, in dem, getrieben durch große dedizierte Teleskope und Instrumente, das Datenaufkommen in den letzten 10 Jahren nahezu exponentiell gewachsen ist. Um die auf diesen Surveys basierenden wissenschaftlichen Fragestellungen weiter verfolgen zu können, sind neue, effektive Analysemethoden notwendig. Es sollen Algorithmen und Datenstrukturen entwickelt und analysiert werden, die eine progressive Exploration sehr großer und hochdimensionaler Datenmengen effizient unterstützen. Ein erstes Ziel ist die Untersuchung massearmer Strukturen, da sie Rückschlüsse auf die Granularität der Masseverteilung (speziell Dunkler Materie) erlauben, die bisher nicht durch Beobachtungen erfasst werden konnte. Speziell für eine solche Suche und Analyse massearmer Strukturen, deren Kontrast zum Vorder- und Hintergrund naturgemäß schwach ist, erwarten wir durch den Einsatz neuer Methoden signifikante Vorteile gegenüber klassischen Verfahren der Stellarstatistik.

VORGEHENSWEISE UND METHODE

Es steht die Konzeption und (prototypische) Umsetzung von Verfahren zum Auffinden von Strukturen in multidimensionalen Datenräumen im Fokus. Als ein erster Schritt soll die Suche nach den von Kugelsternhaufen induzierten *Streams* betrachtet werden. Diese *Streams* mit geringer Masse sind von besonderem Interesse, da sie besonders empfindlich auf Störungen durch vorbeiziehende „Dunkle-Materie-Klumpen“ reagieren.

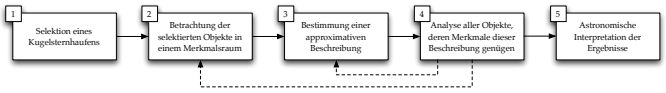

Figure 1. Vorgehensweise bei der Datenanalyse zum Auffinden von *Streams*. Die dargestellte Abfolge der Schritte 2–4 geschieht mehrstufig und parallel für mehrere mögliche (Teil-)Merkmalsräume und wird durch Rückkopplungsschleifen (gestrichelt gezeichnet) unterstützt.

Die zur Detektion und Charakterisierung von *Streams* vorgesehene Vorgehensweise ist in Abbildung 1 schematisch angegeben. Die Selektion eines Kugelsternhaufens erfolgt im ersten Schritt auf Basis der Überdichte hinsichtlich der Merkmale (ra, dec), d.h. in einem zweidimensionalen Merkmalsraum. Basierend auf dem Stand der Forschung, legen wir nun als Arbeitshypothese zu Grunde, dass die Elemente des *Streams* in einem gewissen Merkmalsraum die gleichen Eigenschaften aufweisen wie die Elemente des Kugelsternhaufens. Der zweite Schritt unseres geplanten Vorgehens wird daher darin bestehen, Merkmalsräume zu untersuchen. In jedem dieser Merkmalsräume sollen die Elemente des Kugelsternhaufens auf *Cluster*-Bildung hin untersucht werden. In einem dritten Schritt soll jeder der so gefundenen *Cluster* möglichst kompakt repräsentiert werden. Mit Hilfe dieser Repräsentation werden dann alle Objekte des Katalogs selektiert, die eine vergleichbare Art und Dichte an Merkmalen aufweisen wie die Elemente des Kugelsternhaufens und somit eine Kandidatenmenge für Elemente eines *Streams* bilden. Die Repräsentation der Merkmals-*Cluster* soll zunächst sehr grob approximiert und dann in den folgenden Iterationen immer genauer der Form des *Clusters* angepasst werden. So sollen schnell approximative Lösungen geliefert und für genauere Repräsentationen und mehr zur Verfügung stehende Rechenzeit eine Lösung höherer Güte erzielt werden. Eine Abfolge möglicher Verfeinerungsschritte ist in Abbildung 2 angegeben.

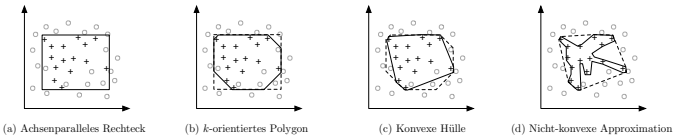

Figure 2. Abfolge feiner werdender, ineinander enthaltenen Approximationen. Die zu approximierenden Punkte im Merkmalsraum sind mit ‘+’ markiert, Punkte, die zu Rauschen korrespondieren, sind mit ‘○’ markiert. Zu den Approximationen (b)–(d) ist die voran gehende Approximation gestrichelt angegeben.

Ein wesentlicher Untersuchungsgegenstand ist daher auch die Abwägung der beiden gegenläufigen Komplexitätsmaße. Der vierte Schritt besteht darin, die so erhaltenen Objekte mit Methoden der Astronomie darauf hin zu untersuchen, ob es sich tatsächlich um Elemente eines durch den Kugelsternhaufens induzierten *Streams* handelt. Im letzten Schritt erfolgt die astronomischen Interpretation der Ergebnisse. Um die Qualität der gefundenen Lösungen bewerten zu können und somit eine (halb-)automatische Exploration des Suchraumes auf der Basis unterschiedlicher geometrischer Beschreibungen von Strukturen zu unterstützen, ist die Definition entsprechender Gütekriterien notwendig. Neben der rein numerischen Bewertung findet in der Astronomie bei der Exploration und Bewertung von Daten häufig eine visuelle Analyse in physikalisch motivierten oder historisch etablierten niedrigdimensionalen (Teil-)Merkmalsräumen statt.

Dieses Vorgehen soll als Ausgangspunkt genutzt werden, um die Güte der geometriebasierten Selektion im Suchraum zu bewerten.

VERWANDTE ARBEITEN

Die Analyse von Daten astronomischer Kataloge ist bislang vorrangig mit Methoden des maschinellen Lernens betrieben worden. In [1] stellen Ball und Brunner in einer sehr umfangreichen Arbeit summarisch die wesentlichen verfolgten Ansätze vor und beurteilen sie bezüglich ihrer Vor- und Nachteile. Als ein zentrales Ergebnis dieser Zusammenstellung ist festzuhalten, dass die Verfahren bislang nicht formal hinsichtlich ihrer Skalierbarkeit für große Datenmengen untersucht wurden. Progressive Verfahren sind in der Vergangenheit vorrangig im Bereich der Datenbanken entwickelt worden, beispielsweise, um frühzeitig Ergebnisse von Verbundoperationen zu erhalten [2], oder um Paretooptimale Punkte zu berechnen [3]. Diese Verfahren wurden jedoch weder bezüglich der klassischen Berechnungskomplexität noch im *two-level disk*-Modell formal analysiert, experimentelle Evaluationen sind nur für kleine Datenmengen ($N \approx 10^6$) bekannt.

VORLÄUFIGE ERGEBNISSE

Die Grundlage zur geometrischen Exploration astronomischer Daten ist durch die Entwicklung eines Werkzeugs geschaffen worden. Dieses bietet die Möglichkeit, astronomische Daten einzulesen und in einem parametrisierbaren Merkmalsraum Zusammenhangskomponenten (ZHK) zu erzeugen. Die ZHK werden auf Basis symmetrischer Adjazenzen bezüglich der *k*-nächsten Nachbarn jedes Datenpunktes gebildet. Anschließend wird eine geometrische Repräsentation für jede ZHK berechnet, die für die Selektion aller Objekte des Katalogs verwendet werden soll. Zusätzlich wurde bereits die Möglichkeit einer visuellen Inspektion realisiert, welche die berechnete Partitionierung durch einen zwei- oder dreidimensionalen Graphen und der berechneten geometrischen Repräsentation darstellt. Somit wird eine rudimentäre Bewertung des beschriebenen ersten Ansatzes zugelassen. Um die visuelle Analyse großer Datensätze zu ermöglichen, wurde bereits ein geometrischer Ansatz zur Erstellung einer repräsentativen Ausdünnung des Datensatzes realisiert.

WEITERE SCHRITTE

Mittelfristig sollen weitere Algorithmen und Methoden entworfen und in das bestehenden Werkzeug integriert werden, um die geometrische Exploration des Datenraums zu verbessern. Hierfür müssen zunächst die Merkmale identifiziert werden, die die intrinsischen Eigenschaften der zu untersuchenden Objekte möglichst kompakt wiedergeben. Da in vielen Fällen die Dimensionalität des Datensatzes erheblichen Einfluss auf die Berechnungskomplexität geometriebasierter Algorithmen hat, ist das Erhalten eines möglichst niedrigdimensionalen Datenraums wünschenswert, es wird aber eine Abwägung zwischen Dimensionalität (und somit Effizienz der Verarbeitung) und Beschreibungsgüte erfolgen müssen. Anschließend müssen die genannten gegenläufigen Kriterien Selektivität und Beschreibungskomplexität bzw. Auswertungskomplexität theoretisch und praktisch gegeneinander abgewogen werden. Zuletzt soll die rein visuelle Bewertung der ermittelten Ergebnisse als Ausgangspunkt genutzt werden, um die Güte der geometriebasierten Selektion im Suchraum zu bewerten.

REFERENCES

1. Nicholas M. Ball and Robert J. Brunner. International Journal of Modern Physics D,19(7) 1049-1106. 2010. Data Mining and Machine Learning in Astronomy.

2. Jens-Peter Dittrich and Berhard Seeger and David Scott Taylor and Peter Widmayer. Progressive Merge-Join: A Generic and Non-Blocking Sort-Based Join Algorithm. VLDB 2002: Proceedings of the 28th International Conference on Very Large Data Bases, 299–310. 2002.

3. Dimitris Papadias and Yufei Tao and Greg Fu and Bernhard Seeger. Progressive Skyline Computation in Database Systems. In ACM Transactions on Database Systems, 30(1) 41–82. 2005.

SLA Calculus

Sebastian Vastag
Informatik LS4, TU Dortmund
Sebastian.vastag@udo.edu
Prof. Peter Buchholz

PROBLEMBESCHREIBUNG UND FORSCHUNGSFRAGE

Einschätzung und Überprüfung von Leistungsvorgaben für service-orientierte Architekturen mittels eines analytischen Verfahres.

Für Webservices und Cloud-Computing werden Systemkomponenten als Dienstleistungen von externen Anbietern eingekauft, diese garantieren ihre Dienstqualität in Form von Service Level Agreements (SLAs). Neben funktionalen Anforderungen sind auch nichtfunktionale Eigenschaften wie Geschwindigkeit und Zuverlässigkeit in SLAs spezifiziert. Bei komponierten Systems soll anhand der gegebenen SLAs obere und untere Schranken für die nichtfunktionalen Eigenschaften des Gesamtsystems abgeschätzt werden. Dazu soll das analytische Werkzeug des Network Calculus auf ein Systemmodell mit SLAs angewendet werden.

Eine modellbasierte Leistungsbestimmung für komponierte Systeme kann helfen, bei Anbietern für Dienstleistungen im Cloud-Computing eine gute Auswahl basierend auf möglicher Leistung und Kosten zu erreichen.

In meiner Forschung arbeite ich an Erweiterungen zum Network Calculus [1]. Insbesondere möchte ich das Wissen über von Kunden akzeptierte Wartezeiten für Webservices mit in das analytische Werkzeug des Calculus integrieren. Besonders spannend ist hier, wie dieses Wissen den Aufbau von parallelisierten Systemen vereinfachen kann.

VORGEHENSWEISE UND METHODE

Die Forschung zum SLA Calculus ist weitestgehend theoretisch und wird durch Auswertungen von Simulationen unterstützt.

Die erreichten Ergebnisse können mit anderen Verfahren aus der Systemtheorie verglichen werden, hier bietet sich insbesondere die Warteschlangentheorie an.

Der Network Calculus ist ein mathematisch fundiertes Verfahren und somit durch Beweise fundiert. Viele Ergebnisse lassen sich direkt auf Service Level Agreements übertragen, eigene Ansätze zur Beschreibung von Wartezeiten im System werden mathematisch bewiesen.

Eine bereits etablierte Methode zur Systembeschreibung ist die Warteschlangentheorie. Sie liefert auf analytischen Weg Mittelwerte zu Leistungsmaßen, dies lässt jedoch keine Aussage über Maxima oder Minima zu. Eine weitere Möglichkeit wäre die Simulation, hier ist es schwer seltene Ereignisse in die Ergebnisse mit einfließen zu lassen.

VERWANDTE ARBEITEN

Eine der wichtigsten Arbeiten im Bereich des Network Calculus ist der Transfer der Methodik auf Eingebettete Systeme [2], dies zeigt die Wandlungsfähigkeit des Kalküls. Den Transfer zwischen einer realen Messung an einen System und den Eingabedaten zum Network Calculus wurde erst sehr spät durch ein statistisches Verfahren ermöglicht [3]. Die vollständige Berechnung der notwendigen Systemgeschwindigkeit durch Last und maximale Wartezeit im Network Calculus wurde als Zwischenergebnis in [4] beschrieben.

Meine Forschung überträgt die Methodik auf ein vollkommen neues Anwendungsgebiet im Bereich der service-orientierten Systeme. Als Novum werden hier Wartezeiten nicht nur als Ergebnis der Möglichen Berechnungen betrachtet, sondern auch als Eingabe. Bisher wurde dies im Kalkül nicht ausgenutzt.

VORLÄUFIGE ERGEBNISSE

Es zeigt sich, dass nichtfunktionale Eigenschaften in Systemen, insbesondere die zu Latenz und Wartezeit, sehr gut durch Funktionen im Network Calculus beschreibbar sind. Dies wesentlich genauer als die Anwendung eines Mittelwertes. Sind die Funktionen zu mehreren Komponenten bekannt lassen sich die begrenzenden Funktionen des Gesamtsystems ableiten.

Dies war im Network Calculus generell schon immer für Angaben zu Last und Arbeitsgeschwindigkeit eines Systems möglich, neu ist dies auch für die Wartezeit. Diese drei Faktoren sind untereinander anhängig, bisherige Zwischenergebnisse zeigen, dass auch beim SLA Calculus der Zusammenhang in einzelnen Systemteilen weiter besteht.

Die weitere Forschung ist darauf ausgerichtet, diese Ergebnisse auch bei seriell und parallel geschaltete Systemkomponenten nachweisen zu können.

WEITERE SCHRITTE

Der SLA Calculus soll auch zur Validierung von SLAs in Simulationsmodellen von service-orientierten Architekturen zu Anwendung kommen. Dazu wird gerade ein am Lehrstuhl vorhandener Simulator für verteilte Softwaresysteme erweitert. Dies ist ein letzter Schritt zur Fertigstellung der Dissertation, er ist zugleich Anwendungsfall und Möglichkeit die Güte der mit dem Calculus berechneten Schranken zu testen.

REFERENCES

[1] Boudec, J.-Y. L., and Thiran, P. Netwark Calculus - A Theory of Deterministic Queu-ing Systems for the Internet, vol. 4 of LNCS. Springer Verlag, May 2004.

[2] Thiele, L., Chakraborty, S., Gries, M., and Künzli, S. A framework for evaluating de-sign tradeoffs in packet processing architectures. In Proceedings of the 39th annual Design Automation Conference (2002), ACM, pp. 880–885.

[3] Undheim, A., Jiang, Y., and Emstad, P. Network calculus approach to router modeling with external measurements. In Communications and Networking in China, 2007. CHINACOM'07. Second International Conference on (2007), IEEE, pp. 276–280.

[4] Fidler, M., and Recker, S. Conjugate network calculus: A dual approach applying the legendre transform. Computer Networks 50, 8 (2006), 1026–1039.

Mikroprotokolle in verdeckten Netzwerkkanälen

Steffen Wendzel
Fakultät für Mathematik und Informatik, FernUniversität in Hagen
Email: steffenwendzel@gmx.de
Betreuer der Doktorarbeit: Prof. Dr. Jörg Keller

PROBLEMBESCHREIBUNG UND FORSCHUNGSFRAGE

Verdeckte Netzwerkkanäle sind von Entwicklern nicht vorgesehene Kommunikationsoptionen in Computernetzen, die einerseits Sicherheitsrichtlinien umgehen und andererseits einen kaum detektierbaren Informationsaustausch erlauben. Mit Kryptografie wird Kommunikation zwar geheim gehalten, bleibt allerdings sichtbar – verdeckte Kanäle kommen hingegen dann zum Einsatz, wenn die Sichtbarkeit einer Kommunikation keine Option darstellt. Anwendung finden können die entwickelten Verfahren etwa bei Journalisten, die in überwachten Netzwerken unbemerkt regimekritische Informationen übertragen müssen. Ebenso könnten Oppositionelle in Ländern mit staatlicher Zensur eine sicherere Kommunikation auf Basis verdeckter Kanäle durchführen.

Mikroprotokolle innerhalb verdeckter Kanäle dienen der Steuerung und Erweiterbarkeit solcher Netzwerkkanäle. Durch dynamische Verfahren innerhalb solcher Mikroprotokolle können Anwender derselben mobil agieren, dynamische Protokollwechsel durchführen und Daten schneller übertragen, als ohne Mikroprotokolle.

Für die zukünftige Anwendung in öffentlichen Netzen, deren Überwachungsverfahren ständiger Weiterentwicklung unterliegen, ist die Kombination von bisherigen Lösungen der Verschlüsselung und neuen (durch Mikroprotokolle gesteuerten) verdeckten Kanäle aussichtsreich.

VORGEHENSWEISE UND METHODE

Die Mikroprotokoll-Ansätze werden durch Software-Implementierungen verifiziert und innerhalb realer Netzwerke analysiert. Störungen gegenüber der verdeckten Kommunikation werden durch den eigentlichen Netzwerktraffic eingeführt und jede Testimplementierung muss diesen Störungen standhalten. Zusätzliche Traffic-Recordings (etwa von Uplinks mit hohem Durchsatz) kommen ebenfalls zum Einsatz. Dieses Vorgehen ist notwendig um die verschiedenen Bedingungen, unter denen Mikroprotokolle operieren müssen, überprüfen zu können. Alternativ könnte eine reine Simulation zum Einsatz kommen, die aber gegenüber realem Traffic und realen Traffic-Recordings weniger aussagekräftig ist.

VERWANDTE ARBEITEN

Die einzige Arbeit im Bereich der Mikroprotokolle für verdeckte Kanäle wurde von Ray und Mishra in [2] veröffentlicht, wobei ein Mikroprotokoll entwickelt wurde, welches bereits durch speichersparsamere Methoden dieser Doktorarbeit effizienter gestaltet werden konnte. Ein Verfahren zur Berechnung der für verdeckte Kanäle verwendbaren Protokolle wurde von Yarochkin et. al. in [1] vorgestellt, wobei auch dieses Verfahren durch die Betrachtung von einzelnen Protokoll-Bestandteilen anstelle ganzer Protokolle ihm Rahmen des Doktorats verbessert wurde.

VORLÄUFIGE ERGEBNISSE

Es konnte gezeigt werden, dass dynamische Protokollwechsel in verdeckten Kanälen möglich sind und die Detektierbarkeit des Kanals verringern können. Zudem konnte der Speicherbedarf von Kanälen reduziert und die mobile Verwendung derselben durch Multi-Protokollabsprachen zwischen Peers ermöglicht werden.

Durch den berechneten Vergleich der Datengrößen mit vorhandenen Mikroprotokollen und die Umsetzung einer Proof-of-Concept-Implementierung konnte die Wirksamkeit und Funktionalität der bisherigen Ergebnisse bestätigt werden.

Weitere Forschung wird den Bereich der kooperativen verdeckten Kanäle in Netzwerken fokussieren. Dabei wird das gemeinsame Benutzen vorhandener Infrastruktur für verdeckte Kanäle im Vordergrund stehen, sodass ganze Anwendergruppen profitieren können. Dabei ist es von Bedeutung, die Risiken solcher Kooperationen gering zu halten. Diese Risiken bestehen primär darin, dass involvierte Partner die Existenz verdeckte Kanäle (oder Infrastruktur-Informationen) an Dritte melden können.

WEITERE SCHRITTE

In Arbeit befindet sich derzeit die Entwicklung kooperativer verdeckter Kanäle. Die Ansätze für diesen Kontext müssen finalisiert und durch eine weitere Implementierung verifiziert werden. Weiterhin muss die Detektierbarkeit des Verhaltens von Mikroprotokollen überprüft werden, weshalb externe Expertise hinsichtlich Network Monitoring von Bedeutung sein wird.

REFERENCES

1. F. V. Yarochkin, S.-Y. Dai et. al.: Towards Adaptive Covert Communication System, In Proc. PRDC, pp. 153-159, 2008.

2. B. Ray, S. Mishra: A Protocol for Building Secure and Reliable Covert Channel, In Proc. Sixth Annual Conference on Privacy, Security and Trust (PST 2008), pp. 246-253, 2008.

Geometrieverarbeitung für die virtuelle Realisierung produktionstechnischer Prozesse

Thomas Wiederkehr
Fakultät Informatik, TU Dortmund
thomas.wiederkehr@tu-dortmund.de
Prof. Dr. Heinrich Müller

PROBLEMBESCHREIBUNG UND FORSCHUNGSFRAGE

Thermisches Spritzen ist ein Prozess, bei dem geschmolzene, mittels einer Gasströmung beschleunigte Metallpartikel auf eine Bauteiloberfläche treffen, dort erstarren und so sukzessiv eine Beschichtung auf der Oberfläche aufbauen. Zur Beschichtung eines Bauteils ist die Bestimmung der Prozessparameter sowie insbesondere einer Roboterbahn, welche die Spritzpistole über die Bauteiloberfläche steuert, erforderlich. Derzeit werden diese Parameter meist durch kosten- und zeitintensive Trial-and-Error-Methoden auf Basis praktischer Versuche bestimmt. Ziel ist daher die Entwicklung einer effizienten Computersimulation zur Bestimmung des Schichtauftrags für thermische Spritzprozesse, welche insbesondere auch im Rahmen einer automatisierten Optimierung der Spritzparameter bzw. der Roboterbahn verwendet werden kann. Da die Simulation dabei sehr häufig aufgerufen wird, ist eine geringe Rechenzeit von hoher Bedeutung. Zur effizienten Implementierung werden daher Methoden aus dem Bereich des GPU-Computings verwendet

VORGEHENSWEISE UND METHODE

Zur Entwicklung eines effizienten Simulationsalgorithmus wurde der Spritzprozess abstrahiert und Parallelen zu Algorithmen in einer (OpenGL) Grafikpipeline identifiziert (Sichtbarkeitsberechnung, Projektion,...). Durch diese Abbildung der Simulation auf die Grafikpipeline wird eine hohe Ausführungsgeschwindigkeit angestrebt, da mit der GPU dedizierte Hardware zur Ausführung der Algorithmen bereitsteht. Die Berechnung der in einem Zeitschritt der Simulation an einem bestimmten Punkt auf der Bauteiloberfläche abgelagerten Materialmenge wird dabei auf Basis geometrischer Gegebenheiten (wie z.B. des Einschlagwinkels) sowie einer experimentell für diesen Prozess ermittelten zweidimensionalen Look-Up-Tabelle in einem Shaderprogramm berechnet. Da die Simulationsergebnisse in hohem Maße von der experimentell ermittelten Look-Up-Tabelle abhängen, wurden hierfür umfangreiche Testreihen mit wechselnden Parametern sowie Versuchswiederholungen mit gleichbleibenden Parametern durchgeführt. Dabei wurden verschiedene unerwünschte Einflussgrößen (beispielsweise Prozessfluktuationen) identifiziert und durch eine algorithmische Nachbearbeitung abgeschwächt (Filter, Surface Fittings). Zur Validierung sind Versuche mit unterschiedlichen Bauteilen durchgeführt und mithilfe des entwickelten Softwaresystems simuliert worden. Die Validierung der Simulationsergebnisse wird anhand des Vergleichs der simulativ ermittelten und der experimentell erzeugten Schichtdicken durchgeführt.

VERWANDTE ARBEITEN

Es gibt nur wenige Arbeiten auf dem Gebiet der Simulation thermischer Spritzprozesse, welche sich mit der Beschichtung ganzer Bauteile befassen. Von der analytischen Materialablagerungsformulierung von [1] wurden einige geometrische Einflussparameter abgeleitet. Eine der wenigen echt dreidimensionalen Simulationen auf Bauteilebene stellt [2] dar, jedoch werden hier aufgrund des hohen Aufwandes Vereinfachungen des Ablagerungsmodells vorgenommen (Rotationssymmetrie). Arbeiten mit Fokus auf GPU-Implementierung und Geschwindigkeit sind keine bekannt.

VORLÄUFIGE ERGEBNISSE

Die Umsetzung der Simulation in OpenGL/GLSL ist sehr schnell. Ferner hat sich die experimentelle Ermittelung der Look-Up-Tabelle als relativ aufwändig erwiesen, wobei störende Einflussgrößen wie z.B. Prozessfluktuationen nicht vollständig auszuschließen sind. Des Weiteren ist noch unklar inwieweit auftretende Abweichungen zwischen Simulation und Experiment auf Prozessschwankungen zurückzuführen sind.

WEITERE SCHRITTE

Zur Prüfung der Ursache der Abweichungen werden weitere Experimente analysiert. Es fehlen jedoch noch genaue Daten zur zu erwartenden Prozesstabilität und -genauigkeit.

REFERENCES

1. Duncan, S.; Jones, P.; Wellstead, P. A frequency domain approach to determining the path separation for spray coating, IEEE Transactions on Automation Sc. and Eng., 2(3), 233-239, 2005.
2. Hansbo, A.; Nylen, P. Models for the simulation of spray deposition and robot motion optimization in thermal spraying of rotating objects, Surf. & Coatings Techn., 122, 191-201. 1999.