



DAGSTUHL REPORTS

Volume 1, Issue 4, April 2011

Plan Recognition (Dagstuhl Seminar 11141) <i>Robert P. Goldman, Christopher W. Geib, Henry Kautz, and Tamim Asfour</i>	1
Innovations for Shape Analysis: Models and Algorithms (Dagstuhl Seminar 11142) <i>Michael Breuß, Alfred M. Bruckstein, and Petros Maragos</i>	23
Formal Methods in Molecular Biology (Dagstuhl Seminar 11151) <i>Rainer Breitling, Adelinde M. Uhrmacher, Frank J. Bruggeman, and Corrado Priami</i>	41
Challenges in Document Mining (Dagstuhl Seminar 11171) <i>Hamish Cunningham, Norbert Fuhr, and Benno M. Stein</i>	65
Artificial Immune Systems (Dagstuhl Seminar 11172) <i>Emma Hart, Thomas Jansen, and Jon Timmis</i>	100

ISSN 2192-5283

Published online and open access by

Schloss Dagstuhl – Leibniz-Zentrum für Informatik GmbH, Dagstuhl Publishing, Saarbrücken/Wadern, Germany.

Online available at <http://www.dagstuhl.de/dagrep>

Publication date

August, 2011

Bibliographic information published by the Deutsche Nationalbibliothek

The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available in the Internet at <http://dnb.d-nb.de>.

License

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivs 3.0 Unported license: CC-BY-NC-ND.



In brief, this license authorizes each and everybody to share (to copy, distribute and transmit) the work under the following conditions, without impairing or restricting the authors' moral rights:

- Attribution: The work must be attributed to its authors.
- Noncommercial: The work may not be used for commercial purposes.
- No derivation: It is not allowed to alter or transform this work.

The copyright is retained by the corresponding authors.

Aims and Scope

The periodical *Dagstuhl Reports* documents the program and the results of Dagstuhl Seminars and Dagstuhl Perspectives Workshops.

In principal, for each Dagstuhl Seminar or Dagstuhl Perspectives Workshop a report is published that contains the following:

- an executive summary of the seminar program and the fundamental results,
 - an overview of the talks given during the seminar (summarized as talk abstracts), and
 - summaries from working groups (if applicable).
- This basic framework can be extended by suitable contributions that are related to the program of the seminar, e.g. summaries from panel discussions or open problem sessions.

Editorial Board

- Susanne Albers
- Bernd Becker
- Karsten Berns
- Stephan Diehl
- Hannes Hartenstein
- Frank Leymann
- Stephan Merz
- Bernhard Nebel
- Han La Poutré
- Bernt Schiele
- Nicole Schweikardt
- Raimund Seidel
- Gerhard Weikum
- Reinhard Wilhelm (*Editor-in-Chief*)

Editorial Office

Marc Herbstritt (*Managing Editor*)

Jutka Gasirowski (*Editorial Assistance*)

Thomas Schillo (*Technical Assistance*)

Contact

Schloss Dagstuhl – Leibniz-Zentrum für Informatik
Dagstuhl Reports, Editorial Office
Oktavie-Allee, 66687 Wadern, Germany
reports@dagstuhl.de

Digital Object Identifier: 10.4230/DagRep.1.4.i

www.dagstuhl.de/dagrep

Plan Recognition

Edited by

Robert P. Goldman¹, Christopher W. Geib², Henry Kautz³, and
Tamim Asfour⁴

- 1 SIFT, LLC – Minneapolis, US, rpgoldman@sift.net
- 2 University of Edinburgh, GB, cgeib@inf.ed.ac.uk
- 3 University of Rochester, US, kautz@cs.rochester.edu
- 4 KIT – Karlsruher Institut für Technologie, DE, asfour@kit.edu

Abstract

This Dagstuhl seminar brought together researchers with a wide range of interests and backgrounds related to plan and activity recognition. It featured a substantial set of longer tutorials on aspects of plan and activity recognition, and related topics and useful methods, as a way of establishing a common vocabulary and shared basis of understanding. Building on this shared understanding, individual researchers presented talks about their work in the area. There were also panel discussions which addressed questions about how to best foster progress in the field — specifically how to improve our ability to compare different plan and activity recognition algorithms — and address the question of whether to assume rationality in the modeled agents (a question that is of great concern in many fields at this time). This report presents a summary of the talks and discussions at the seminar.

Seminar 03.–08. April, 2011 – www.dagstuhl.de/11141

1998 ACM Subject Classification I.2.m Artificial Intelligence, Miscellaneous

Keywords and phrases Artificial intelligence, plan recognition, intent recognition, activity recognition.

Digital Object Identifier 10.4230/DagRep.1.4.1


1 Executive Summary

Robert P. Goldman

Christopher W. Geib

Tamim Asfour

Henry Kautz

License  Creative Commons BY-NC-ND 3.0 Unported license
© Robert P. Goldman, Christopher W. Geib, Tamim Asfour, and Henry Kautz

Plan recognition, activity recognition, and intent recognition all involve making inferences about other actors from observations of their behavior, i.e., their interaction with the environment and with each other. The observed actors may be software agents, robots, or humans. This synergistic area of research combines and unifies techniques from user modeling, machine vision, intelligent user interfaces, human/computer interaction, autonomous and multi-agent systems, natural language understanding, and machine learning. It plays a crucial role in a wide variety of applications including:

- assistive technology
- software assistants



Except where otherwise noted, content of this report is licensed under a Creative Commons BY-NC-ND 3.0 Unported license

Plan Recognition, *Dagstuhl Reports*, Vol. 1, Issue 4, pp. 1–22

Editors: Robert P. Goldman, Christopher W. Geib, Henry Kautz, and Tamim Asfour



Dagstuhl Reports

Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

- computer and network security
- behavior recognition
- coordination in robots and software agents
- e-commerce and collaborative filtering

This Dagstuhl seminar brought together researchers with a wide range of interests and backgrounds related to plan and activity recognition. It featured a substantial set of longer tutorials on aspects of plan and activity recognition, and related topics and useful methods, as a way of establishing a common vocabulary and shared basis of understanding. These were:

- Plan recognition and discourse;
- Plan recognition and psychology;
- Probabilistic methods;
- Plan recognition and learning;
- Grammatical methods and
- Planning and plan recognition.

The common ground constructed by these tutorials provided a basis that individual researchers could build upon when sharing their specific interests and developments.

One challenge to progress in plan recognition is that there has not been a shared agreement about what constitutes plan recognition: what are its inputs and outputs, and what constitutes a good answer. In particular, this has inhibited progress because it is difficult to clearly compare new work in plan recognition with preceding work (quantitative comparisons are almost impossible), there is a paucity of shared data sets, etc. Coming into the seminar, the organizing committee proposed that the field might be improved by the introduction of a plan recognition competition, modeled on competitions in AI planning (the International Planning Competition), SAT solving, etc. Discussions at the seminar concluded that it would be premature to introduce such a competition at this time. Participants felt that a more productive use of community resources would be to develop a shared repository of plan and activity recognition data sets. A number of participants volunteered to provide their data sets, and there has been movement towards establishing a common public repository.

Plan Recognition: background

The earliest work in plan recognition was rule-based; researchers attempted to come up with inference rules that would capture the nature of plan recognition. However without an underlying formal model these rule sets are difficult to maintain and do not scale well.

In 1986, Kautz and Allen (K&A) published an article, “Generalized Plan Recognition” [7] that framed much of the work in plan recognition to date. K&A defined the problem of plan recognition as the problem of identifying a minimal set of *top-level actions* sufficient to explain the set of observed actions. Plans were represented in a plan graph, with top-level actions as root nodes and expansions of these actions into unordered sets of child actions representing plan decomposition. To a first approximation, the problem of plan recognition was then a problem of graph covering. K&A formalized this view of plan recognition in terms of McCarthy’s circumscription. Kautz [6] presented an approximate implementation of this approach that recast the problem as one of computing vertex covers of the plan graph.

A number of early plan recognition systems used techniques such as rule-based systems [9], vertex covering, etc. Such techniques are not able to take into account differences in the *a priori* likelihood of different goals. Observing an agent going to the airport, this algorithm

views “air travel,” and “terrorist attack” as equally likely explanations, since they explain (cover) the observations equally well.

To the best of our knowledge, Charniak was the first to argue that plan recognition was best understood as a specific case of the general problem of *abduction*, or reasoning to the best explanation [3, e.g.,]. Charniak and Goldman (C&G) [2] argued that, viewing plan recognition as abduction, it could best be done as Bayesian (probabilistic) inference. Bayesian inference supports the preference for minimal explanations, in the case of equally likely hypotheses, but also correctly handles explanations of the same complexity but different likelihoods. For example, if a set of observations could be equally well explained by two hypotheses, theft and bragging being one, and theft alone being the other, simple probability theory (with some minor assumptions), will tell us that the simpler hypothesis is the more likely one. On the other hand, if as above, the two hypotheses were “air travel” and “terrorist attack,” and each explained the observations equally well, then the prior probabilities will dominate, and air travel will be seen to be the most likely explanation. There have been many similar approaches to the problem, based on cost minimization, etc.

Another broad area of attack on the problem of plan recognition has been to reformulate it as a parsing problem [10, e.g.,]. Parsing-based approaches to plan recognition promise greater efficiency than other approaches, but at the cost of making strong assumptions about the ordering of plan steps. The major problem with parsing as a model of plan recognition is that it does not treat partially-ordered plans or interleaved plans well. Approaches that use statistical parsing [11, e.g.,] combine parsing and Bayesian approaches.

Finally, there has been a large amount of very promising work done using variations of Hidden Markov Models (HMMs) [1], techniques that came to prominence in signal processing applications, including speech recognition. These approaches offer many of the efficiency advantages of parsing approaches, but with the additional advantages of incorporating likelihood information and of supporting machine learning to automatically acquire their plan models. Standard HMMs seem to be insufficiently expressive to capture planful behavior, but a number of researchers have extended them to hierarchical formulations, that capture more complicated intentions. Conditional Random Fields [8], dynamic Bayes nets, and other probabilistic models have also been used.

Much of this latter work has been done under the rubric of *activity recognition*. The early work in this area very carefully chose the term *activity* or *behavior recognition* to distinguish it from plan recognition. The distinction to be made between activity recognition and plan recognition is the difference between recognizing a single (possibly complex) activity and recognizing the relationships between a set of such activities that result in a complete plan. Much of the work on activity recognition can be seen as discretizing a sequence of possibly noisy and intermittent low-level sensor readings into coherent actions that could be treated as inputs to a plan recognition system.

Several researchers have been interested in using plan recognition to improve team coordination [4, 5]. That is, if agents in a team can recognize what their teammates are doing, then they can better cooperate and coordinate. They may also be able to learn something about their shared environment. For example, a member of a military squad who sees a teammate ducking for cover may infer that there is a threat, so that it also takes precautions.

References

- 1 H. Bui and S. Venkatesh and G. A. W. West. Policy recognition in the abstract Hidden Markov Model. In *Journal of Artificial Intelligence Research*, vol. 17, pages 451–499, 2002.

- 2 Eugene Charniak and Robert P. Goldman. A Bayesian model of plan recognition. *Artificial Intelligence*, 64(1):53–79, November 1993.
- 3 Eugene Charniak and Drew McDermott. *Introduction to Artificial Intelligence*. Addison Wesley, Reading, MA, 1985.
- 4 Marcus J. Huber, Edmund H. Durfee and Michael P. Wellman. The Automated mapping of plans for plan recognition. In *Proceedings of the National Conference on Artificial Intelligence*, pages 344–351, 1994.
- 5 Gal A. Kaminka, D.V. Pynadath, and M. Tambe. Monitoring teams by overhearing: a multi-agent plan-recognition approach. In *Journal of Artificial Intelligence Research*, vol. 17, pages 83–135, 2002.
- 6 Henry Kautz. *A Formal Theory of Plan Recognition*. Technical Report, Department of Computer Science, University of Rochester, May 1987.
- 7 Henry Kautz and James F. Allen. Generalized plan recognition. In *Proceedings of the National Conference on Artificial Intelligence*, pages 32–38, 1986.
- 8 Lin Liao, Dieter Fox and Henry Kautz. Learning and inferring transportation routines. In *Proceedings of the National Conference on Artificial Intelligence*, 2004.
- 9 C.F. Schmidt and N.S. Sridharan and J.L. Goodson. The plan recognition problem: an intersection of psychology and artificial intelligence. In *Artificial Intelligence*, vol. 11, pages 45–83, 1978.
- 10 Marc Vilain. Getting serious about parsing plans: a grammatical analysis of plan recognition. In *Proceedings of the National Conference on Artificial Intelligence (AAAI-90)*, pages 190–197, 1990.
- 11 David V. Pynadath and Michael P. Wellman. Accounting for context in plan recognition with application to traffic monitoring. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, pages 472–481, 1995.

2 Table of Contents

Executive Summary

Robert P. Goldman, Christopher W. Geib, Tamim Asfour, and Henry Kautz 1

Overview of Talks

From Motion to Text and Back for Humanoid Robots

Tamim Asfour 7

Bayesian Theory of Mind: Modeling Joint Belief-Desire Inference

Chris L. Baker 7

Eliciting Plan Recognition Cues by Provoking Opponents in RTS Games

Francis Bisson 8

Knowledge-rich Plans – How they Enable Explanation, Recognition, and Repair

Susanne Biundo 8

Thinking about Evaluation and Corpora for Plan Recognition

Nate Blaylock 9

“The gist of the matter”: On plan understanding, behaviour prediction, and referring expressions

Michael Brenner 9

AbRA: An Abductive, Rationalizing Agent for Plan Recognition

Will Bridewell 9

Plan Recognition for User-Adaptive Interaction

Cristina Conati 10

Activity Recognition for a Knowledge Worker Assistant

Thomas Dietterich 10

Tutorial: Plan Recognition via Inverse Reinforcement Learning

Thomas Dietterich 10

Plan Recognition and Collaborative Assistants

George Ferguson 11

Planning and Plan Recognition

Hector Geffner 11

Grammatical Methods for Plan Recognition

Christopher W. Geib 12

Plan recognition: a historical survey, part I

Robert P. Goldman 12

Behavior Recognition and Demonstration for Human/Robot Cooperation

Tetsunari Inamura 12

Plan recognition challenges in real-time strategy games

Frodoald Kabanza 13

Survey of Probabilistic Activity and Plan Recognition

Henry A. Kautz 13

Mobile Intention Recognition And Spatially Constrained Grammars

Peter Kiefer 14

Probabilistic Plan Recognition <i>Kathryn B. Laskey</i>	14
Plan Recognition Using Multi-Entity Bayesian Networks and PR-OWL <i>Kathryn B. Laskey</i>	15
Plan Recognition in/with Agent Programming Languages <i>Yves Lesperance</i>	15
Yappr: From LL parsing to plan recognition <i>John Maraist</i>	16
No More Plan Libraries – The Case for a Structureless World <i>David Pattison</i>	16
Modeling Theory of Mind as Plan Recognition <i>David Pynadath</i>	16
Intentions in Collaboration: Insights from Meaning <i>Matthew Stone</i>	17
Assuming the Human’s Cognitive State as Basis for Assistant System Initiative <i>Ruben Strenzke</i>	18
Coupling Plan Recognition with Plan Repair for Real-Time Opponent Modeling <i>Gita Reese Sukthankar</i>	19
Efficient Hybrid Algorithms for Plan Recognition and Detection of Suspicious and Anomalous Behavior <i>Dorit Zilberbrand</i>	19
Panel Discussions	
A Plan recognition competition? <i>Christopher W. Geib</i>	20
Rational versus fallible agents <i>Matthew Stone</i>	20
Invited Talks	21
Participants	22

3 Overview of Talks

3.1 From Motion to Text and Back for Humanoid Robots

Tamim Asfour (KIT – Karlsruhe Institute of Technology, DE)

License © ⓘ ⊖ Creative Commons BY-NC-ND 3.0 Unported license
© Tamim Asfour

Joint work of Asfour, Tamim; Dillmann, Ruediger

URL <http://his.anthropomatik.kit.edu/english/65.php>

Semantic representations are a prerequisite for the development of cognitive capabilities and understanding in robots as well as for cooperation, interaction and communication with humans. Building such representations from sensorimotor experience rely on organizing the system’s sensorimotor experience to provide data structures which can be used at different levels of the systems hierarchy and breaks through the gap between sensorimotor level and symbolic level.

In this talk, we present our recent work on building humanoid robots able to act, interact and autonomously acquire knowledge in the real world. Results are presented towards the implementation of integrated 24/7 humanoid robots able to 1) perform complex grasping and manipulation tasks in a kitchen environment 2) autonomously acquire object knowledge through active visual and haptic exploration and 3) learn actions from human observation and imitate them in goal-directed manner.

Further, we discuss how a motion library can be built from observation of human demonstration and how the elements of such library can be represented and enriched by additional constraints such as objects involved in an action, forces applied to an object and agents involved in interaction and cooperation tasks.

The resulting data structures, together methods of natural language processing will facilitate the link between sensorimotor experience and linguistic representations.

3.2 Bayesian Theory of Mind: Modeling Joint Belief-Desire Inference

Chris L. Baker (MIT - Cambridge, US)

License © ⓘ ⊖ Creative Commons BY-NC-ND 3.0 Unported license
© Chris L. Baker


Joint work of Baker, Chris L.; Saxe, Rebecca R.; Tenenbaum, Joshua B.

Main reference Baker, C.L., Saxe, R.R., Tenenbaum, J.B., “Bayesian Theory of Mind,” Proceedings of the Thirty-Second Annual Conference of the Cognitive Science Society, to appear.

We present a computational framework for understanding Theory of Mind (ToM): the human capacity to make joint inferences about the beliefs and desires underlying the observed actions of other agents. Bayesian ToM (BToM) formalizes the concept of intentional agency at the heart of ToM as a partially observable Markov decision process (POMDP), and performs Bayesian inference over this structured model to reconstruct an agent’s joint belief state and reward function given observations of its behavior in some environmental context. We test the BToM framework by collecting people’s joint inferences of agents’ desires and beliefs about unobserved aspects of the environment in response to stimuli of agents moving in simple spatial scenarios. BToM performs substantially better than two simpler variants: one in which desires are inferred without reference to an agents’ beliefs, and another in which beliefs are inferred without reference to the agent’s dynamic observations in the environment.

3.3 Eliciting Plan Recognition Cues by Provoking Opponents in RTS Games

Francis Bisson (Université de Sherbrooke, CA)

License  Creative Commons BY-NC-ND 3.0 Unported license
© Francis Bisson

Joint work of Bisson, Francis; Kabanza, Froduald; Benaskeur, Abder Rezak; Irandoust, Hengameh


Main reference Bisson, F., Kabanza, F., Benaskeur, A. and Irandoust, H., “Provoking Opponents to Facilitate the Recognition of their Intentions,” Proceedings of the AAAI Student Abstract and Poster Program, 2011.

URL <http://planiart.usherbrooke.ca/bisson/papers/aaai2011-poster.pdf>

For agents evolving in adversarial environments such as RTS games, it is necessary to be able to recognize the goals of their opponents in the environment. However, most adversarial plan recognizers rely on a passive observation of the opponents, gathering and analyzing cues related to their goals. In contrast, in this talk I will present preliminary results for a plan recognition approach that can provoke the opponents in order to observe their reactions, and use the resulting cues to disambiguate the current set of hypotheses on their goals.

3.4 Knowledge-rich Plans – How they Enable Explanation, Recognition, and Repair

Susanne Biundo (Universität Ulm, DE)


License  Creative Commons BY-NC-ND 3.0 Unported license
© Susanne Biundo

Hybrid planning combines the traditional planning paradigms of hierarchical task network (HTN) and partial-order causal-link (POCL) planning. The resulting systems are able to use predefined standard solutions like in pure HTN planning, but can also develop (parts of) a plan from scratch or modify a default solution in cases where the initial state deviates from the presumed standard. This flexibility makes hybrid planning particularly well suited for real-world applications.

Based on a completely declarative description of actions, tasks, and solution methods, hybrid planning smoothly integrates reasoning about procedural knowledge and causalities and allows for the generation of knowledge-rich plans of action. The information those plans comprise includes causal dependencies between actions on both abstract and primitive levels as well as information about their hierarchical and temporal relationships. By making use of this information, as well as of the underlying declarative domain models, capabilities like the generation of courses of action on various abstraction levels, the stable repair of failed plans, plan recognition, and the explanation of different solutions for a given planning problem can be implemented by advanced automated reasoning techniques.

3.5 Thinking about Evaluation and Corpora for Plan Recognition

Nate Blaylock (IHMC – Pensacola, US)


License  Creative Commons BY-NC-ND 3.0 Unported license
© Nate Blaylock

Joint work of Blaylock, Nate; Allen, James

Lately, many fields of AI have benefited from labeled corpora as common resources for training and evaluating performance. I will discuss some of the issues of creating corpora for plan recognition and argue that this would be a worthwhile investment for our community. I will also discuss a range of metrics for evaluating the performance of plan recognizers.

3.6 “The gist of the matter”: On plan understanding, behaviour prediction, and referring expressions

Michael Brenner (Universität Freiburg, DE)


License  Creative Commons BY-NC-ND 3.0 Unported license
© Michael Brenner

The talk discussed the problem of understanding the purpose of a plan. I argued that this is more than recognising the goal state it tries to achieve, but rather a characterisation of the plan in relation to "deficits" in (and other constraints on) the initial state. Such a characterisation (called a "gist") can be described indepently of the specific initial state by making use of referring expressions, similarly to their use in natural-language processing.

The talk then discussed cost measures for gists and some initial ideas for recognising them given an observed plan.

3.7 AbRA: An Abductive, Rationalizing Agent for Plan Recognition

Will Bridewell (Stanford University, US)

License  Creative Commons BY-NC-ND 3.0 Unported license
© Will Bridewell

Joint work of Bridewell, Will; Langley, Pat


Main reference Bridewell, W., & Langley, P., “A computational account of everyday abductive inference,” Proceedings of the 33rd Annual Meeting of the Cognitive Science Society, 2011.

Plan recognition is a naturally abductive task. That is, looking at the actions of an agent, one must make assumptions about its underlying plan. AbRA is a novel, logic-based system for carrying out socially aware, abductive inference.

This system emphasizes cyclic, online operation, incremental extension of explanations, a shifting focus of attention, and a data-driven inference mechanism. Guided by local coherence, AbRA constructs an explanation/plan that ties the observations into a plausible, although not necessarily correct or even optimally rational, story. Here I provide an intuitive description of the system, report preliminary results on a complex plan recognition domain, and plot our current research trajectory.

3.8 Plan Recognition for User-Adaptive Interaction


Cristina Conati (University of British Columbia – Vancouver, CA)

License  Creative Commons BY-NC-ND 3.0 Unported license
© Cristina Conati

I will first give some examples of how we use plan/goal/activity recognition in user-adaptive interactive systems. I will then introduce two directions we are exploring to improve the accuracy and usability of user-adaptive interaction: (i) using eye-gaze information to inform plan recognition; (ii) Explaining to the user aspects of the system’s reasoning to increase user trust in the system’s adaptive interventions

3.9 Activity Recognition for a Knowledge Worker Assistant

Thomas Dietterich (Oregon State University, US)

License  Creative Commons BY-NC-ND 3.0 Unported license
© Thomas Dietterich


Joint work of Dietterich, Thomas; Shen, Jianqiang; Bao, Xinlong; Keiser, Victoria; Bui, Hung
Main reference Dietterich, T. G., Bao, X., Keiser, V., Shen, J., “Machine Learning Methods for High Level Cyber Situation Awareness,” In Jajodia, S., Liu, P., Swarup, V. and Wang, C. (Eds.) , “*Cyber Situation Awareness*”. Springer, 2009, pp. 227–247.

URL <http://web.engr.oregonstate.edu/tgd/publications/csa-dietterich-bao-keiser-shen.pdf>

Knowledge workers execute hundreds of simple digital workflows in a typical work week. We will describe three forms of activity recognition that seek to assist knowledge workers with these workflows. The first is the TaskTracer Project Predictor, which attempts to infer which project the user is working on based on observed desktop activity. The second is the TaskTracer Folder Predictor, which predicts which folder the user wishes to access when opening or saving a file. Experimental studies show that Folder Predictor reduces by 50% desired folder. The third is a method for discovering and recognizing workflow executions as part of an effort to provide proactive assistance to desktop knowledge workers.

3.10 Tutorial: Plan Recognition via Inverse Reinforcement Learning

Thomas Dietterich (Oregon State University, US)

License  Creative Commons BY-NC-ND 3.0 Unported license
© Thomas Dietterich





Reinforcement learning methods seek an optimal policy for an unknown Markov Decision Process by interacting with that process to learn the transition and reward functions. Inverse Reinforcement learning is given the transition function and the optimal (or expert) policy and seeks to find the reward function. More generally, Inverse RL can be viewed as attempting to infer the goals underlying observed behavior. A closely-related task is to infer a expert’s policy from demonstrations.

The components of an MDP (reward function, policy, value function, state visitation probabilities) are inter-related, and Inverse RL methods can be categorized based on the primary component that they attempt to learn. Given observed behavior, there are multiple reward functions and value functions consistent with it (even asymptotically), which makes direct attempts to learn these components ill-defined. In contrast, methods that seek to

directly learn the policy or the state visitation probabilities appear to be more successful, because these are uniquely specified by observed behavior (at least asymptotically). This tutorial surveys several methods for learning reward functions, state visitation probabilities, and policies and concludes that learning state visitation probabilities is the most promising approach.

3.11 Plan Recognition and Collaborative Assistants





George Ferguson (University of Rochester, US)

License     Creative Commons BY-NC-ND 3.0 Unported license
© George Ferguson

This talk discusses the roles of plan recognition in the design and implementation of collaborative assistants—intelligent systems that interact naturally to help people solve problems. Two key roles are identified: (1) the disambiguation of natural language input to support mixed-initiative interaction and learning from demonstration, and (2) tracking user performance during execution to support both mixed-initiative interaction, task- and context-sensitive help, and overt instruction or teaching. These roles are illustrated with examples from systems we have implemented in the past. We also describe a new research thrust based on combining natural language description with low-level sensor data for learning models of real-world tasks performed by humans.

3.12 Planning and Plan Recognition

Hector Geffner (Universidad Pompeu Fabra – Barcelona, ES)

License     Creative Commons BY-NC-ND 3.0 Unported license
© Hector Geffner

Joint work of Geffner, Hector; Ramirez, Miquel

Main reference M. Ramirez and H. Geffner, “Probabilistic Plan Recognition using off-the-shelf Classical Planners,” Proc. AAAI-2010.

URL <http://www.dtic.upf.edu/~hgeffner>

Plan recognition is like planning in reverse: while in planning the goal is given and a plan is sought; in plan recognition, part of a plan is given, and the goal and complete plan are sought. Until recently, however, plan recognition has been addressed using methods which are not related to planning such as parsing algorithms, Bayesian network procedures, and specialized methods.


In almost all cases, the space of possible plans or activities to be recognized is assumed to be given by a suitable library or set of policies.

Recently, an approach that does not require the use of a plan library and which uses planning technology, as been developed by Baker, Saxe, and Tenenbaum, on the one hand, and by Ramirez and Geffner, on the other. In this approach, the plan recognition problem is mapped into a collection of planning problems that can be solved with off-shelf-planners. The posterior distribution over the possible goals given the observation is inferred from basic probability laws and costs derived from the use of a planner. The approach has been used to perform planning recognition over classical planning models, Markov Decision Processes (MDPs), and Partially Observable MDPs (POMDPs).

In this invited talk, I review the relevant ideas from AI Planning and their use for formulating and solving the plan recognition problem.

3.13 Grammatical Methods for Plan Recognition


Christopher W. Geib (University of Edinburgh, GB)

License  Creative Commons BY-NC-ND 3.0 Unported license
© Christopher W. Geib

This talk is an overview of prior and current work on viewing plan recognition as a parsing task given a formal grammar and a sequence of observations. It covers work in using: regular, context free, probabilistic state-dependent, plan tree, tree adjoining, and combinatorial categorial grammars.

3.14 Plan recognition: a historical survey, part I

Robert P. Goldman (SIFT – Minneapolis, US)


License  Creative Commons BY-NC-ND 3.0 Unported license
© Robert P. Goldman

The seminar opened with a historical survey of plan recognition. We present a taxonomy of plan recognition problems, including the conventional distinction between keyhole, intended, and adversarial plan recognition, but touching on other dimensions such as fallible versus ideal agents, complete versus partial observability, open versus closed worlds, static versus evolving sets of intentions, and expressiveness of plan representation. After outlining the dimensions of plan representation, we proceeded to review methods used in the early history of plan recognition. We began with early techniques, based on rule-based systems, then moved on to discuss the formalization of the field, based on Kautz and Allen’s generalized theory of plan recognition, and Vilain’s parsing-based complexity analysis of the theory. We discussed systems inspired by this work, including techniques based on parsing and on minimal graph cover.

We concluded the first part of this talk with a discussion of techniques based on Bayes networks.

3.15 Behavior Recognition and Demonstration for Human/Robot Cooperation

Tetsunari Inamura (NII – Tokyo, JP)

License  Creative Commons BY-NC-ND 3.0 Unported license
© Tetsunari Inamura

Main reference Tetsunari Inamura, Keisuke Okuno, “Robotic Motion Coach: Effect of Motion Emphasis and Verbal Expression for Imitation Learning,” Proc. 3rd International Conference on Cognitive Neurodynamics, p.186, 2011.

Behavior Recognition and Demonstration for Human-Robot Cooperation In this talk, development of a robotic coaching system is discussed. recognition and reproduction of human’s whole body motion patterns are focused towards establishment of natural human-robot

cooperation through demonstration of motion and speech conversation. A robotic coaching system should be a target application because the robot should recognize user's motion, demonstrate modified motion according to user's level of skill, and generate advice with verbal expression. Abstract of motion pattern using HMMs and a phase space are proposed. Using the phase space, motion emphasis and generation of verbal expression are integrated. A robotic simulator platform is also introduced towards a basis of evaluation of plan recognition for human-robot interaction application.

References

- 1 Tetsunari Inamura and Keusuke Okuno. *Robotic Motion Coach: Effect of Motion Emphasis and Verbal Expression for Imitation Learning*. The 3rd International Conference on Cognitive Neurodynamics, p.186, 2011.

3.16 Plan recognition challenges in real-time strategy games

Froduald Kabanza (Université de Sherbrooke, CA)

License © © ⊖ Creative Commons BY-NC-ND 3.0 Unported license
© Froduald Kabanza

Main reference Kabanza, F., Bellefeuille, P., Bisson, F., Benaskeur, A., and Irandoust, H., "Opponent Behaviour Recognition for Real-Time Strategy Games," Proc. of AAAI Workshop on Plan, Activity and Intent Recognition (PAIR), 2010.

URL <http://planiart.usherbrooke.ca/kabanza/publications/10/pair10-opponent.pdf>

In real-time strategy (RTS) games, players recruit and manoeuvre army units in order to defeat their opponents. The victory condition may vary from one game or scenario to another, but it usually involves destroying some or all of the opponent's assets. A key component of the player's situation awareness in this context is the recognition of his opponent's intent and plans. This presentation covers some of the main challenges posed by the intent and plan recognition problems in RTS games and sketch the main building blocks of a conceptual plan recognition method geared towards addressing these challenges. The method is still a concept in the early development stage, and the presentation will be aimed at stimulating a discussion and encouraging the audience to comment on it rather than demonstrating its effectiveness. The RTS domain is used for concrete scenarios, but the fundamental intent and plan recognition problems that we are addressing remain relevant to other adversarial domains.

3.17 Survey of Probabilistic Activity and Plan Recognition

Henry A. Kautz (University of Rochester, US)

License © © ⊖ Creative Commons BY-NC-ND 3.0 Unported license
© Henry A. Kautz


We provide an overview of probabilistic plan recognition methods. These include:

- HMM
- Layered HMM
- Dynamic Bayesian Networks
- Stochastic Grammars
- Conditional Random Fields
- Relational Markov Model

- Markov Logic
- Bayesian Inverse Planning
- Inverse Reinforcement Learning
- N-Gram Models

3.18 Mobile Intention Recognition And Spatially Constrained Grammars

Peter Kiefer (Universität Bamberg, DE)

License  Creative Commons BY-NC-ND 3.0 Unported license
© Peter Kiefer

Main reference Peter Kiefer, “The Mobile Intention Recognition Problem And An Approach Based On Spatially-Constrained Grammars,” PhD Thesis, to appear 2011.


Mobile intention recognition differs from the general plan and intention recognition problem by the availability of spatial context information for each input behavior. This talk proposes to use the specific properties of spatial context, such as continuity and hierarchies, for the disambiguation of mobile behavior sequences.

Most current approaches for interpreting mobile behavior focus on activity recognition, not on (high-level) intentions. This talk argues that formal grammars, enhanced with spatial information, are well-suited for representing high-level intentions. Formal grammars make expressiveness properties explicit, are cognitively comprehensible, and allow for easy geographic portability - a requirement crucial in mobile assistance domains.

Many mobile assistance domains require us to represent behaviors and intentions with formal grammars of higher expressiveness than context-free grammars. This talk proposes to enhance the mildly context-sensitive Tree-Adjoining Grammar formalism, well-known in natural language processing, with spatial constraints, yielding in a spatial grammar specifically useful to express the Visit-/Revisit-pattern frequently occurring in mobile assistance.

3.19 Probabilistic Plan Recognition

Kathryn B. Laskey (George Mason University – Fairfax, US)


License  Creative Commons BY-NC-ND 3.0 Unported license
© Kathryn B. Laskey

Plan recognition is naturally viewed as a problem in inference under uncertainty. From observations of an agent’s actions (or effects actions), a plan recognition system attempts to infer the agent’s goal and explain the actions in terms of a plan for achieving the goal. Typically, there are multiple explanations for any sequence of actions. Probability is a natural approach to weighing the relative plausibility of alternative explanations.

Attractive features of probability include its strong theoretical foundation, its unified view of inference and learning, and its practical success in a growing body of applications. On the other hand, probabilistic inference and learning are NP-hard, and achieving sufficiently expressive yet tractable representations is a major challenge. This talk provides an overview of major probabilistic representations used for plan recognition, describes common exact and approximate inference methods, and identifies research challenges.

3.20 Plan Recognition Using Multi-Entity Bayesian Networks and PR-OWL

Kathryn B. Laskey (George Mason University – Fairfax, US)


License  Creative Commons BY-NC-ND 3.0 Unported license
© Kathryn B. Laskey

Joint work of Carvalho, Rommel N.; Costa, Paulo C. G.; Laskey, Kathryn B.; and Chang, KuoChu
Main reference Carvalho, Rommel N.; Costa, Paulo C. G.; Laskey, Kathryn B.; Chang, KuoChu, “PROGNOS: Predictive Situational Awareness with Probabilistic Ontologies,” Proceedings of the Thirteenth International Conference of the Society of Information Fusion (FUSION 2010).
URL http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5711970

Increasingly expressive languages are emerging for representing and reasoning with probability. Multi-Entity Bayesian Networks (MEBN) is a first-order language for specifying probabilistic knowledge bases as parameterized fragments of Bayesian networks. MEBN fragments (MFrag) can be instantiated and combined to form arbitrarily complex graphical probability models. An MFrag represents probabilistic relationships among a conceptually meaningful group of uncertain hypotheses. The PR-OWL probabilistic ontology language, based on MEBN, extends OWL to allow expression of uncertainty about attributes and relations. An example is given of a MEBN theory for maritime domain awareness and its implementation as a PR-OWL probabilistic ontology.

3.21 Plan Recognition in/with Agent Programming Languages

Yves Lespérance (York University – Toronto, CA)

License  Creative Commons BY-NC-ND 3.0 Unported license
© Yves Lesperance

Joint work of Lesperance, Yves; Goultiaeva, Alexandra
Main reference Goultiaeva, A. and Lespérance, Y., “Incremental Plan Recognition in an Agent Programming Framework,” In Working Notes of the AAAI 2007 Workshop on Plan, Activity, and Intent Recognition (PAIR’07), Vancouver, BC, July, 2007.
URL <http://www.cse.yorku.ca/~lesperan/papers/PAIR07.pdf>

In the talk, I discuss how agent programming languages can be used for specifying plans for plan recognition, and also how plan recognition capabilities could be usefully added to such languages. I focus on the ConGolog agent programming language, based on the situation calculus. I review an account of plan recognition for this setting [1], where ConGolog plan libraries are used. This provides a very expressive language for specifying plans.


It supports several forms of nondeterminism and allows sketchy plan templates to be specified. Also it is closed under union, intersection, and complementation, so one can specify the set of runs that are part of the plan in a completely compositional way. I discuss how the account can be simplified by restricting attention to situation-determined programs, where the remaining program after a partial execution is uniquely determined [2].

References

- 1 Goultiaeva, A. and Lespérance, Y., “Incremental Plan Recognition in an Agent Programming Framework.” In *Working Notes of the AAAI 2007 Workshop on Plan, Activity, and Intent Recognition (PAIR’07)*, Vancouver, BC, July, 2007.
- 2 De Giacomo, G., Lespérance, Y., and Muise, C., “Agent Supervision in Situation-Determined ConGolog.” To appear in *Working Notes of the 9th International Workshop on Nonmonotonic Reasoning, Action and Change (NRAC-2011)*, Barcelona, Spain, July, 2011.

3.22 Yappr: From LL parsing to plan recognition

John Maraist (SIFT – Minneapolis, US)


License  Creative Commons BY-NC-ND 3.0 Unported license
© John Maraist

Joint work of Geib, Christopher W.; Goldman, Robert P.; Maraist, John
Main reference Christopher W. Geib, John Maraist, Robert P. Goldman, “A New Probabilistic Plan Recognition Algorithm Based on String Rewriting,” Proc. of the 18th International Conference on Automated Planning and Scheduling (ICAPS-2008), Sydney, Australia, 2008.

We present the probabilistic HTN plan recognition algorithm Yappr by evolution from its motivating classical parsing algorithm. We begin with a simple stack-based automaton for nondeterministic LL parsing, and identify the three refinements which produce Yappr: precompilation of plans for efficient, deterministic retrieval; replacement of the parser stack with a graph allowing multiple application points; and maintenance of multiple explanations rather than a single parse. We conclude with a look forward at the advantages and disadvantages of moving from an LL-based to an LR- based approach.

3.23 No More Plan Libraries – The Case for a Structureless World

David Pattison (The University of Strathclyde – Glasgow, GB)

License  Creative Commons BY-NC-ND 3.0 Unported license
© David Pattison


Joint work of Pattison, David; Long, Derek

Plan Libraries have always gone hand-in-hand with Plan Recognition. Having a concise set of possible plans to map to an agent’s observed actions allows for reliable goal recognition, next-action prediction, intermediate states and further prediction and analysis.

The problem is that these libraries don’t exist in the real world. Construction of a plan/goal library by hand is a labour-intensive process with a highly bespoke output. In the past 10 years work has moved towards the generation of these libraries at runtime, but computer-generated plans can never be truly perfect, with invalid or unwanted entries an unavoidable side-effect. In this talk I will discuss my own work on Goal Recognition as a Planning problem as an example of such a model, and the need to move away from a library-based standard in recognition.

3.24 Modeling Theory of Mind as Plan Recognition

David Pynadath (University of Southern California – Marina del Rey, US)

License  Creative Commons BY-NC-ND 3.0 Unported license
© David Pynadath


Human social interaction relies on our ability to model each other as (mostly) rational actors. Despite the uncertainty we may have of another’s intentions and subjective beliefs, our theory of mind provides valuable leverage that we exploit whenever possible. Thus, multiagent modeling of social situations can benefit from a computational implementation of theory of mind. I present one such implementation where an agent reuses its own decision-theoretic planning to generate expectations about the behavior of others. By combining its uncertain beliefs about the possible mental states of others with a planning mechanism, it becomes

straightforward to recast this problem as plan recognition. Inverting the planning process generates abductive reasoning that an agent can use to update its beliefs about others as it observes their behavior. While such recursive beliefs can become prohibitively complex as the number of agents increases, I also show that the decision-theoretic context gives each agent a utility-based metric for deciding what it can safely ignore about everyone else.

We have implemented these algorithms within a social simulation framework, PsychSim, that has supported simulations of various scenarios, including bilateral negotiation, language and cultural training, and urban stabilization operations.

3.25 Intentions in Collaboration: Insights from Meaning

Matthew Stone (Rutgers University – Piscataway, US)

License  Creative Commons BY-NC-ND 3.0 Unported license
© Matthew Stone

Conversation is one of many cases where we want to attribute intentions to agents exhibiting improvised, fluid, expert strategic behavior. Understanding an utterance, on the received view, is just recognizing the speaker's communicative intention. But does this make sense? Any action could implicitly prepare for an open-ended array of contingencies, reflecting an open-ended array of expectations its agent brings to the situation. And agents may well choose those actions through heuristic problem solving and learned strategies—processes that fit poorly with the intuitive notions of deliberation and commitment used by intention theorists. How can we actually carve out recognizable intentions from such a complex ensemble of factors? Think of recognizing the intention of the Roshambo player you see throw Rock, who might specifically intend to make a throw chosen at random, to play best response against you, or both.

In this talk, I will try to clarify both the received view of meaning as intention and the place of intention recognition in collaborative activity. We have surprisingly strong judgments about what people can and cannot mean with individual utterances, and about how those meanings fit together over the course of a conversation. These judgments motivate a specific kind of collaborative intention: a system of public categories of action, coordinated across agent and teammates, classifying each action based on the content of the mental representation that immediately underpins its performance. Playing Rock fits in such a system, playing at random or playing best response do not. I close by sketching how an implemented agent—one that might not actually meet traditional standards for having individual or shared intentions—could use such categories to pursue its utterances in a meaningful and collaborative way.

3.26 Assuming the Human’s Cognitive State as Basis for Assistant System Initiative

Ruben Strenzke (Unibw – München, DE)

License © © ⊖ Creative Commons BY-NC-ND 3.0 Unported license

© Ruben Strenzke

Joint work of Strenzke, Ruben; Schulte, Axel

Main reference Strenzke, Ruben; Schulte, Axel, “Modeling the Human Operator’s Cognitive Process to Enable Assistant System Decisions,” Proc. of Goal, Activity and Plan Recognition (GAPRec) Workshop in conjunction with International Conference on Automated Planning and Scheduling (ICAPS) 2011.

URL <http://icaps11.icaps-conference.org/proceedings/gaprec/strenzke.pdf>

In the Manned-Unmanned Teaming application a transport helicopter commander is also controlling multiple reconnaissance unmanned aerial vehicles (UAVs) that shall reduce the risk of the transport mission. This human operator is therefore responsible of planning and re-planning a multi-aircraft mission under situation-dependent time pressure. To lower the workload generated thereby, he/she shall be supported by a cognitive assistant system that is designed with respect to the Cooperative Automation and mixed-initiative planning approaches.

In order to decide, whether, when, and in which way to take initiative, the assistant system has to know about the human operator’s goal, assume his/her plan, and evaluate his/her activity. The latter is also necessary in order to estimate the current workload situation.

In our implementation the assistant system assumes the human plan by taking into account the mission order (goal constraints) and the partial or complete plan entered by the human into the system (plan constraints). The assistant system then checks, if this plan is feasible, complete, and compare this plan with what itself has planned automatically. On this basis the assistant system can decide to take initiative to urge the operator to improve plan quality and enforce timely plan execution.

For another task of this human operator (identification of vehicles during route reconnaissance by UAVs) there have been workload estimation experiments based on operator activity, i.e. his/her manual and visual interaction with the system.

The workload estimation can be accomplished by HMMs per operator task and per workload level. In case of workload higher than normal, so-called self-adaptive strategies are observable, which alter the human behavior.

The great challenge remains the combination of methods like used in the two approaches mentioned. This would allow to assume the operator’s cognitive state in a broader context, which makes sense because in the cognitive planning process the goals lead to the plan, the plan leads to choosing action, action leads to behavior, and workload leads to errors induced in the different steps of this process.

3.27 Coupling Plan Recognition with Plan Repair for Real-Time Opponent Modeling

Gita Reese Sukthankar (University of Central Florida – Orlando, US)

License © © ⊖ Creative Commons BY-NC-ND 3.0 Unported license
© Gita Reese Sukthankar

Joint work of Sukthankar, Gita Reese; Laviers, Kennard

Main reference Kennard Laviers, Gita Sukthankar, “A Real-time Opponent Modeling System for Rush Football,” Proceedings of International Joint Conference on Artificial Intelligence, July 2011.

One drawback with using plan recognition in adversarial games is that often players must commit to a plan before it is possible to infer the opponent’s intentions. In such cases, it is valuable to couple plan recognition with plan repair, particularly in multi-agent domains where complete replanning is not computationally feasible. This paper presents a method for learning plan repair policies in real-time using Upper Confidence Bounds for Trees (UCT). We demonstrate how these policies can be coupled with plan recognition in an American football game (Rush 2008) to create an autonomous offensive team capable of responding to unexpected changes in defensive strategy. Our real-time version of UCT learns play modifications that result in a significantly higher average yardage and fewer interceptions than either the baseline game or domain-specific heuristics.

3.28 Efficient Hybrid Algorithms for Plan Recognition and Detection of Suspicious and Anomalous Behavior

Dorit Zilberbrand (Givat Shmuel, IL)


License © © ⊖ Creative Commons BY-NC-ND 3.0 Unported license
© Dorit Zilberbrand

Plan recognition is the process of inferring other agents’ plans and goals based on their observable actions. Modern applications of plan recognition, in particular in surveillance and security raise several challenges. First, a number of key capabilities are missing from all but a handful of plan recognizers: (a) handling complex multi-featured observations; (b) dealing with plan execution duration constraints; (c) handling lossy observations (where an observation is intermittently lost); and (d) handling interleaved plans. Second, essentially all previous work in plan recognition has focused on recognition accuracy itself, with no regard to the use of the information in the recognizing agent. As a result, low-likelihood recognition hypotheses that may imply significant meaning to the observer, are ignored in existing work. In this work we present set of efficient plan recognition algorithms that are capable of handling the variety of features required of realistic recognition tasks. We also present novel efficient algorithms that allow the observer to incorporate her own biases and preferences, in the form of a utility function, into the plan recognition process. This allows choosing recognition hypotheses based on their expected utility to the observer. We call this Utility-based Plan Recognition (UPR). We demonstrate the efficacy of the techniques described above, by applying them to the problem of detecting anomalous and suspicious behavior. The system contains the symbolic plan recognition algorithm, which detects anomalous behavior, and the utility-based plan recognizer which reasons about the expected cost of hypotheses. These two components form a highly efficient hybrid plan recognizer capable of recognizing abnormal and potentially dangerous activities. We evaluate the system with extensive experiments, using real-world and simulated activity data, from a variety of sources.

4 Panel Discussions

4.1 A Plan recognition competition?


Christopher W. Geib (University of Edinburgh – Edinburgh, UK)

License  Creative Commons BY-NC-ND 3.0 Unported license
© Christopher W. Geib

Little of the research published in plan recognition reports on results that can be directly compared to previous research. The community has yet to agreed on standard data sets or benchmark problems that all systems are expected to be evaluated against. In an effort to address this same problem the AI planning research community established the International Planning Competition (IPC). At Dagstuhl, we had a panel discussion to consider if a similar competition would benefit the plan recognition community. The pannel members were Christopher Geib, Hector Geffner, Jerry Hobbs, and Froduald Kabanza. There was lively debate both pro and con, and while there were strong arguments in favor it was not universally agreed that an IPC-style competition would be in the best interests of the plan recognition community. There was significant concern that such a competition could fragment the small and still growing plan recognition community and might unintentionally limit future research directions. That said, it was generally agreed that more efforts should be made, especially by leaders in the community, to share data sets and report work in a way that enabled more directly comparable results. A number of participants in the seminar agreed to make their data sets freely available.

4.2 Rational versus fallible agents

Matthew Stone (Rutgers University – Piscataway, US)

License  Creative Commons BY-NC-ND 3.0 Unported license
© Matthew Stone

The panel on rational versus fallible agents primarily addressed the issues involved in using automated plan recognition approaches to understand human activity. There are several respects in which humans may not be perfect decision makers. For example, as Professor Conati observed, student learners solving problems may apply incorrect approaches or use erroneous facts. She found it crucial to model these mistakes in her systems for plan recognition for intelligent tutoring. Similarly, Professor Stone pointed out that many examples of indirection in conversation seem to rely on the heuristics and biases of human decision making, and cited a number of likely cases from the work of Steven Pinker. His dialogue systems increasingly assume very constrained reasoning on the part of interlocutors. At the same time, however, Chris Baker emphasized that there are many domains where people do exhibit expert behavior which systems need to understand. Assumptions of rationality can be very effective in these domains in making good predictions with simple models and minimal training. Meanwhile, Professor Kautz observed that many learning approaches to plan recognition put the focus on finding reliable patterns of activity, and make few assumptions one way or the other about the rationality of target agents. Looking forward, the panel recommended that researchers aim to factor out assumptions about agents from their algorithms wherever possible, so that the community can focus on techniques that generalize across diverse agent populations and tasks. It is also important to evaluate

this dimension of plan recognition systems, to understand where and when assumptions of rationality are violated, and what effects such cases have both on the performance of plan recognition algorithms and the contribution of plan recognition to broader measures of system performance.

5 Invited Talks

Invited talks have their abstracts (where available) in the main body of the report.

- Plan recognition and discourse, Jerry Hobbs
- Plan recognition and psychology, Chris Baker
- Probabilistic methods, Kathryn Blackmond Laskey
- Plan recognition and learning, Tom Dietterich
- Grammatical methods, Christopher W. Geib
- Planning and plan recognition, Hector Geffner

Participants

- Tamim Asfour
KIT – Karlsruhe Institute of
Technology, DE
- Chris L. Baker
MIT – Cambridge, US
- Francis Bisson
Université de Sherbrooke, CA
- Susanne Biundo
Universität Ulm, DE
- Nate Blaylock
IHMC – Pensacola, US
- Michael Brenner
Universität Freiburg, DE
- Will Bridewell
Stanford University, US
- Cristina Conati
University of British Columbia –
Vancouver, CA
- Thomas Dietterich
Oregon State University, US
- George Ferguson
University of Rochester, US
- Hector Geffner
Universidad Pompeu Fabra –
Barcelona, ES
- Christopher W. Geib
University of Edinburgh, GB
- Robert P. Goldman
SIFT – Minneapolis, US
- Jerry Hobbs
University of Southern California
– Marina del Rey, US
- Tetsunari Inamura
NII – Tokyo, JP
- Froduald Kabanza
Université de Sherbrooke, CA
- Henry A. Kautz
University of Rochester, US
- Peter Kiefer
Universität Bamberg, DE
- Kathryn B. Laskey
George Mason University –
Fairfax, US
- Yves Lesperance
York University – Toronto, CA
- John Maraist
SIFT – Minneapolis, US
- David Pattison
The University of Strathclyde –
Glasgow, GB
- David Pynadath
University of Southern California
– Marina del Rey, US
- Matthew Stone
Rutgers Univ. – Piscataway, US
- Ruben Strenzke
Unibw – München, DE
- Gita Reese Sukthankar
University of Central Florida –
Orlando, US
- Dorit Zilberbrand
Givat Shmuel, IL



Innovations for Shape Analysis: Models and Algorithms

Edited by

Michael Breuß¹, Alfred M. Bruckstein², and Petros Maragos³

1 Universität des Saarlandes, DE

2 Technion – Haifa, IL, freddy@cs.technion.ac.il

3 National TU – Athens, GR, maragos@cs.ntua.gr

Abstract

This report documents the program and the results of Dagstuhl Seminar 11142 *Innovations for Shape Analysis: Models and Algorithms*, taking place April 3-8 in 2011. The focus of the seminar was to discuss modern and emerging topics in shape analysis by researchers from different scientific communities, as there is no conference specifically devoted to this field.

Seminar 03.–08. April, 2011 – www.dagstuhl.de/11142

1998 ACM Subject Classification I.4 Image Processing and Computer Vision

Keywords and phrases Shape analysis, mathematical morphology, shape reconstruction, numerical computing, level set methods, fast marching methods

Digital Object Identifier 10.4230/DagRep.1.4.23


Edited in cooperation with Silvano Galliani

1 Executive Summary

Michael Breuß

Alfred M. Bruckstein

Petros Maragos

License  Creative Commons BY-NC-ND 3.0 Unported license
© Michael Breuß, Alfred M. Bruckstein, and Petros Maragos

The notion of *shape* is fundamental in image processing and computer vision, as the shape of objects allows the semantic interpretation of image contents. This is also known from human vision, as humans can recognise characteristic objects solely from their shapes. By *shape analysis* one denotes models and algorithms for detection and processing of shapes in images. It is at the heart of a lot of applications in sciences and engineering.

While the visual quality of an image benefits from a large number of pixels and a high resolution of details, the same assertion does not generally hold for important shape information. As a simple example, when refining an image of $N = n \times n$ pixels towards a higher resolution, the number of pixels in the contour of an object grows just linearly with n , but the image size N grows quadratically. This *linear scaling property* of shape descriptors is an attractive feature of shape analysis methods when considering large images.

Thanks to technological progress made within the last decade one can nowadays acquire high-resolution images with relatively inexpensive, common devices. An important example are digital cameras that allow to acquire images of several megapixels. The size of the image files has grown accordingly. As the trend of developing even more accurate and inexpensive acquisition devices will certainly continue in the next years, the linear scaling property of shape descriptors makes shape analysis methods an even more useful tool in image processing than in the past. However, there are also substantial *new challenges* in



Except where otherwise noted, content of this report is licensed under a Creative Commons BY-NC-ND 3.0 Unported license

Innovations for Shape Analysis: Models and Algorithms, *Dagstuhl Reports*, Vol. 1, Issue 4, pp. 23–40

Editors: Michael Breuß, Alfred M. Bruckstein, and Petros Maragos



Dagstuhl Reports

Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

shape analysis: Concerning algorithms that allow the processing of the arising large data sets in acceptable time, and with respect to adequate shape analysis models that allow for an efficient algorithmic formulation.

The purpose of this seminar was to meet these challenges by bringing together researchers that are engaged in recent and upcoming developments in shape analysis models and numerical computing. As an example, the field of differential geometry has grown to be important for shape analysis during the last years, while a field like deformable shape modelling just begins to influence shape analysis methods. On the algorithmical side, there are many recent innovations that can be important for shape analysis. As examples, let us mention new broadly applicable, efficient fast marching schemes, or graph-based iterative algorithms. The individual areas in shape analysis and numerical computing share an interest in the described techniques. However, modelling is seen as a hot topic in computer science, while numerical computing is often seen as a mathematical domain. Also the various areas within shape analysis research can benefit from the discussion of models and methods that are modern in their respective fields.

The purpose of bringing together researchers from different disciplines was to explore the benefits of a *cross-disciplinary* point of view.

- Researchers in continuous-scale shape analysis brought their knowledge of differential and variational models and the related methods to the meeting.
- Researchers in discrete shape analysis brought to the meeting their knowledge about the latest techniques in graph-based shape analysis, discrete topology and related optimisation methods.
- Researchers in numerical computing brought to the meeting their knowledge of numerical techniques and of numerical analysis.

As the demands in the individual fields of shape analysis are high, the research groups in which the most interesting techniques are under development are quite specialised. Because of this, there is no regular conference or workshop that serves as a meeting place for an exchange of ideas of these groups.

The seminar was conducted in a conference style, where every contributor gave a talk of about 20 to 25 minutes. There was much time for extensive discussions in between the talks and in the evenings, and as documented by the very positive evaluation there was generally a very open and constructive atmosphere. While it is at the moment this report is written very difficult to identify a new fundamental aspect of shape analysis as a result of the workshop, lots of interesting aspects were discussed. As we believe, these will inspire novel developments in both modelling and algorithms.

2 Table of Contents

Executive Summary

<i>Michael Breuß, Alfred M. Bruckstein, and Petros Maragos</i>	23
--	----

Overview of Talks


Geodesic Variations: Algorithms and Applications <i>Fethallah Benmansour</i>	27
Refined Homotopic Thinning Algorithms and Quality Measures for Skeletonisation Methods <i>Michael Breuß</i>	27
Local and Global Diffusion Geometry in Non-rigid Shape Analysis <i>Alex M. Bronstein</i>	28
Intrinsic Symmetry <i>Michael M. Bronstein</i>	28
Video analysis: a Tool Towards Unsupervised Learning of Shapes <i>Thomas Brox</i>	29
From MLS to Overparametrized Nonlocal Variational Methods <i>Alfred M. Bruckstein</i>	29
A Semi-Lagrangian Scheme for the AMSS Model <i>Elisabetta Carlini</i>	29
Modeling Morse Complexes in Arbitrary Dimensions <i>Leila De Floriani</i>	30
Shape Analysis Problems in Practical Applications <i>Stephan Didas</i>	30
Non-rigid Shape Correspondence by Matching Pointwise Surface Descriptors and Metric Structures <i>Anastasia Dubrovina</i>	30
Semi-Lagrangian Schemes for Nonlinear PDEs in Image Processing <i>Maurizio Falcone</i>	31
Geodesic Regression on Shape Manifolds <i>P. Thomas Fletcher</i>	31
Statistical Shape Models in Biomedical Image Segmentation and Visualization of Fuzzy Shapes <i>Hans-Christian Hege</i>	32
Segmentation and Skeletonization on Arbitrary Graphs Using Multiscale Morphology and Active Contours <i>Petros Maragos</i>	32
A PDE approach to Photometric Shape from Shading <i>Roberto Mecca</i>	32
Discrete surface processing and adaptive remeshing <i>Serena Morigi</i>	33

A truly Unsupervised, Non-Parametric Clustering Method <i>Pablo Muse</i>	33
Image-based 3D Modeling via Cheeger Sets <i>Martin Oswald</i>	34
Nested Sphere Statistics of Skeletal Models <i>Stephen M. Pizer</i>	34
Group-valued regularization for motion segmentation of dynamic non-rigid shapes <i>Guy Rosman</i>	35
Variational Models in Shape Space and Links to Continuum Mechanics <i>Martin Rumpf</i>	35
3D Curve Skeleton Computation and Use for Discrete Shape Analysis <i>Gabriella Sanniti Di Baja</i>	35
Incremental Level Set Tracking <i>Nir Sochen</i>	36
Global Minimization for Continuous Multiphase Partitioning Problems Using a Dual Approach and graph cuts algorithms <i>Xue-Cheng Tai</i>	36
Non-Local Ambrosio-Tortorelli and 3-Partite Skeletons <i>Sibel Tari</i>	37
Integrated DEM Construction and Calibration of Hyperspectral Imagery: A Remote Sensing Perspective <i>Christian Woehler</i>	37
Stochastic Diffeomorphic Evolution and Tracking <i>Laurent Younes</i>	38
Shape Analysis for 3D Point Cloud. <i>Hong-Kai Zhao</i>	38
Distance Images and Intermediate-Level Vision <i>Steven W. Zucker</i>	39
Orientation and Anisotropy of Multicomponent Shapes <i>Jovisa Zunic</i>	39
Participants	40

3 Overview of Talks

3.1 Geodesic Variations: Algorithms and Applications

Fethallah Benmansour (EPFL – Lausanne, CH)

License  Creative Commons BY-NC-ND 3.0 Unported license
© Fethallah Benmansour


Joint work of Benmansour, Fethallah; Santambrogio, Filippo; Carlier, Guillaume; Peyré, Gabriel

The computation of geodesic distance and minimal paths is an essential building block for many techniques and applications in computer vision, computer graphics and imaging. In all these fields, it makes sense to consider variational optimization problems that takes into account geodesic distances. This includes, for instance optimal sampling of surfaces, shape optimization, traffic congestion, travel time seismic imaging, surface reconstruction, etc. Several efficient numerical solvers are available for the computation of geodesic distances.

However, the numerical computation of the derivative of this distance with respect to some parameters has been much less investigated. In this talk I will first review the Sub-Gradient Marching Algorithm, which is a numerical procedure to evaluate the derivative of the geodesic distance with respect to the metric on regular grids. I will detail a novel class of algorithms that extend this work to triangulated mesh, and can compute derivative with respect to the metric and to the seeding points. I will show some applications to shape optimization, traffic congestion, and geodesic centroidal Voronoi tessellation.

3.2 Refined Homotopic Thinning Algorithms and Quality Measures for Skeletonisation Methods

Michael Breuß (Universität des Saarlandes, DE)


License  Creative Commons BY-NC-ND 3.0 Unported license
© Michael Breuß

Topological skeletons are shape descriptors that have been applied successfully in practical applications. However, many skeletonisation methods lack accessibility, mainly due to the need for manual parameter adjustment and the shortage of tools for comparative analysis.

In our contribution we address these problems. We propose two new homotopy-preserving thinning algorithms: Flux-ordered adaptive thinning (FOAT) extends existing flux-based thinning methods by a robust automatic parameter adjustment, maximal disc thinning (MDT) combines maximal disc detection on Euclidean distance maps with homotopic thinning. Moreover, we propose distinct quality measures that allow to analyse the properties of skeletonisation algorithms. Tests of the new algorithms and quality assessment tools are conducted on the widely used shape database CE-Shape-1.

3.3 Local and Global Diffusion Geometry in Non-rigid Shape Analysis

Alex M. Bronstein (Tel Aviv University, IL)

License  Creative Commons BY-NC-ND 3.0 Unported license
© Alex M. Bronstein

Diffusion geometry, scale-space analysis, and study of heat propagation on manifolds have recently become a popular tool in data analysis in a variety of applications. In this talk, we will explore the applications of diffusion geometry to the problems of non-rigid shape representation, comparison, and retrieval. We will show that diffusion processes allow defining both local and global geometric structures. Local shape descriptors based on diffusion kernels allow representing shapes as collections of geometric "words" and "expressions" and approaching shape similarity as problems in text search and matching. Global structures are diffusion metrics, insensitive to shape deformations and topological changes. Representing shapes as metric spaces endowed with diffusion distances, we can pose the problem of shape similarity as a comparison of metric spaces using the Gromov-Hausdorff distance. As examples of applications we will show large-scale shape retrieval, correspondence computation, and detection of intrinsic symmetries in non-rigid shapes.

3.4 Intrinsic Symmetry

Michael M. Bronstein (Universität Lugano, CH)

License  Creative Commons BY-NC-ND 3.0 Unported license
© Michael M. Bronstein

Joint work of Bronstein, Michael M.; Raviv, Dan; Bronstein, Alexander; Kimmel, Ron; Hooda, Amit; Horaud, Radu


Symmetry and self-similarity is the cornerstone of Nature, exhibiting itself through the shapes of natural creations and ubiquitous laws of physics. Since many natural objects are symmetric, the absence of symmetry can often be an indication of some anomaly or abnormal behavior. Therefore, detection of asymmetries is important in numerous practical applications, including crystallography, medical imaging, and face recognition, to mention a few. Conversely, the assumption of underlying shape symmetry can facilitate solutions to many problems in shape reconstruction and analysis.

Traditionally, symmetries are described as extrinsic geometric properties of the shape. While being adequate for rigid shapes, such a description is inappropriate for non-rigid ones. Extrinsic symmetry can be broken as a result of shape deformations, while its intrinsic symmetry is preserved.

In this talk, we will present a generalization of symmetries for non-rigid shapes and two different numerical framework for their detection and classification.

3.5 Video analysis: a Tool Towards Unsupervised Learning of Shapes


Thomas Brox (Universität Freiburg, DE)

License  Creative Commons BY-NC-ND 3.0 Unported license
© Thomas Brox

Unsupervised learning requires a grouping step that defines which data belong together. A natural way of grouping in images is the segmentation of objects or parts of objects. While pure bottom-up segmentation from static cues is well known to be ambiguous at the object level, the odds are much better as soon as objects move. I will present a method that uses long term point trajectories based on dense optical flow. Defining pair-wise distances between these trajectories allows to cluster them, which results in temporally consistent segmentations of moving objects in a video shot. In contrast to multi-body factorization, points and even whole objects may appear or disappear during the shot.

3.6 From MLS to Overparametrized Nonlocal Variational Methods


Alfred M. Bruckstein (Technion – Haifa, IL)

License  Creative Commons BY-NC-ND 3.0 Unported license
© Alfred M. Bruckstein

This talk discusses a variational methodology, which involves locally modeling of data from noisy samples, combined with semi-local or global model-parameters regularization. We show that this methodology yields many previously proposed algorithms, from the celebrated moving least squares (MLS) methods to the globally optimal over-parametrization methods recently published for smoothing and optic flow estimation. The unified look at the range of problems and methods previously considered also suggests a wealth of novel global functionals and local modeling possibilities. Moreover, the proposed non-local variational functional provided by this methodology greatly improves the robustness and accuracy compared to previous methods. Therefore the proposed methodology may be viewed as a basis for a general framework for a variety of problem domains in signal and image processing and analysis, such as denoising, adaptive smoothing, reconstruction and segmentation.

3.7 A Semi-Lagrangian Scheme for the AMSS Model

Elisabetta Carlini (University of Rome “Sapienza”, IT)


License  Creative Commons BY-NC-ND 3.0 Unported license
© Elisabetta Carlini

Joint work of Carlini, Elisabetta; Ferretti, Roberto

I will present a Semi-Lagrangian scheme for the approximation of the Mean Curvature Motion power 1/3, better known in image community as Affine Morphological Scale Space (AMSS) model. I will analyse then the properties of the scheme: consistency, monotonicity and convergence. I will finally present some numerical test for image de-noising.

3.8 Modeling Morse Complexes in Arbitrary Dimensions


Leila De Florianì (University of Genova, IT)

License  Creative Commons BY-NC-ND 3.0 Unported license
© Leila De Florianì

Ascending and descending Morse complexes, determined by a scalar field f defined over a manifold M , induce a subdivision of M into regions associated with the critical points of f . In this talk, we present two simplification operators, called removal and contraction, for Morse complexes, which work in arbitrary dimensions. Together with their inverse refinement operators, we show that they form a minimally complete set of atomic operators to create and update Morse complexes on M . We also present a compact dimension-independent graph-based representation for Morse complexes which can be coupled with a discrete representation of the field as a simplicial mesh. We describe the effect of the simplification operators on such representation. Finally, we show some results and discuss current work and future perspectives.

3.9 Shape Analysis Problems in Practical Applications


Stephan Didas (Fraunhofer ITWM – Kaiserslautern, DE)

License  Creative Commons BY-NC-ND 3.0 Unported license
© Stephan Didas

In this talk, we are sketching some shape analysis and characterisation problems motivated by current projects of the image processing department at Fraunhofer ITWM. The goals of these projects include automatic quality assurance in the production line or characterisation of different newly developed materials like fiber-reinforced polymers in material sciences, for example. In quality assurance applications, requirements on the system's response time often severely limit the choice of algorithms useable for the given task. Nevertheless, the presence of texture on the objects under investigation or the aim of allowing for a very general setting for the imaging process necessitate sophisticated methodologies to obtain the desired information from the image data. In material sciences, the measurement techniques are sometimes driven to their extremes (for example, in terms of spatial resolution). Thus the images can only have a very limited quality which makes a reliable analysis more complicated. Therefore, the intention of the presentation is not to present working solutions. We rather point out some (according to our knowledge) open problems from a practical point of view.

3.10 Non-rigid Shape Correspondence by Matching Pointwise Surface Descriptors and Metric Structures

Anastasia Dubrovina (Technion – Haifa, IL)

License  Creative Commons BY-NC-ND 3.0 Unported license
© Anastasia Dubrovina




Finding a correspondence between two non-rigid shapes is one of the cornerstone problems in the field of three-dimensional shape processing. We describe a framework for marker-less non-rigid shape correspondence, based on matching intrinsic invariant surface descriptors, and the

metric structures of the shapes. The matching task is formulated as a quadratic optimization problem that can be used with any type of descriptors and metric, and solved using an integer optimization tool. Further, we show the correspondence ambiguity problem arising when matching intrinsically symmetric shapes using only intrinsic surface properties. We show that when using isometry invariant surface descriptors based on eigendecomposition of the Laplace-Beltrami operator, it is possible to construct distinctive sets of surface descriptors for different possible correspondences. When used in a proper minimization problem, those descriptors allow us to explore number of possible correspondences between two given shapes.

Finally, we describe a hierarchical framework for an efficient solution of the quadratic matching problem.

3.11 Semi-Lagrangian Schemes for Nonlinear PDEs in Image Processing

Maurizio Falcone (University of Rome “Sapienza”, IT)

License     Creative Commons BY-NC-ND 3.0 Unported license
© Maurizio Falcone

The PDE approach to the solution of image processing problems has been rather successful in the last 30 years opening new directions in the field. Typically the PDEs appearing in some classical problems like Shape from Shading, segmentation and filtering are nonlinear and often degenerate so that standard numerical methods for their approximation can not be applied. We describe a class of numerical schemes for first and second order nonlinear PDEs based on semi-Lagrangian (SL) techniques. Those methods have been originally applied for simple advection equations, more recently they have been extended to deal with a wide range of nonlinear possibly degenerate differential problems. The main advantage is that SL schemes have strong stability properties and allow for large time steps still providing accurate solutions. We will give a short review of these properties as well as some applications to some classical image processing problems.

3.12 Geodesic Regression on Shape Manifolds


P. Thomas Fletcher (University of Utah, US)

License     Creative Commons BY-NC-ND 3.0 Unported license
© P. Thomas Fletcher

In this talk I present a regression method, called geodesic regression, for modeling the relationship between a manifold-valued random variable and a real-valued independent parameter. The principle is to fit a geodesic curve, parameterized by the independent parameter, that best fits the data. Error in the model is evaluated as the sum-of-squared geodesic distances from the model to the data, and this provides an intrinsic least squares criterion. Geodesic regression is, in some sense, the simplest parametric model that one could choose, and it provides a direct generalization of linear regression to the manifold setting. I will also present a hypothesis test for determining the significance of the estimated trend. While the method can be generally applied to any form of manifold data, I will show a specific example of analyzing shape changes in the corpus callosum due to age.

3.13 Statistical Shape Models in Biomedical Image Segmentation and Visualization of Fuzzy Shapes

Hans-Christian Hege (ZIB – Berlin, DE)


License  Creative Commons BY-NC-ND 3.0 Unported license
© Hans-Christian Hege

In the first part of the talk a toolset is sketched for creation of 3D statistical shape models (including semi-automatic segmentation, geometry processing for creation of training shapes, flexible establishment of correspondences and PCA in Euclidean space) and segmentation of image data using various methods for adaptation of the model to new image data (including analysis of 1D intensity profiles along normal vectors or omni-directional, formulation of the optimization as MRF and computing the optimum using FastPD). Furthermore, the creation and utilization of articulated statistical shape models for modeling of anatomical joints is addressed.

In the second part of the talk the propagation of uncertainties (errors, noise) in a scalar field to positional uncertainty of level sets is discussed. This includes the role of numerical condition as well as computation of spatial probabilities that a given element (edge, face, ...) of a cell of the sampling grid crosses a level set of a given threshold. Two cases are discussed: uncorrelated random variables with arbitrary distribution and correlated Gaussian random variables. Examples from engineering and climate research illustrate the results.

3.14 Segmentation and Skeletonization on Arbitrary Graphs Using Multiscale Morphology and Active Contours

Petros Maragos (National TU – Athens, GR)


License  Creative Commons BY-NC-ND 3.0 Unported license
© Petros Maragos

Joint work of Maragos, Petros; Kimon Drakopoulos

In this chapter we focus on formulating and implementing on abstract structures such as arbitrary graphs popular methods and techniques developed for image analysis, in particular multiscale morphology and active contours. To this goal we extend existing work on graph morphology to multiscale dilation and erosion and implement them recursively using level sets of functions defined on the graph's nodes. We propose approximations to the calculation of the gradient and the divergence of vector functions defined on graphs and use these approximations to apply the technique of geodesic active contours for object and edge detection. Finally, using these novel ideas, we propose a method for multiscale shape skeletonization on arbitrary graphs.

3.15 A PDE approach to Photometric Shape from Shading

Roberto Mecca (University of Rome “Sapienza”, IT)


License  Creative Commons BY-NC-ND 3.0 Unported license
© Roberto Mecca

The Shape from Shading problem based on a single light source suffers from the convex/concave ambiguity. One possible solution to this limitation is to use two or more light sources,

considering the Shape from Shading Photometric Stereo (SfS-PS) problem. In the talk I will introduce the SfS-PS problem and present some new theoretical results showing uniqueness. In fact, the use of two or more images based on different light sources allows to determine, under some hypotheses, the surface we are observing without additional information other than the light sources direction. I will also describe some numerical schemes based on the semi-lagrangian approximation of the new differential problem, discussing their efficiency and accuracy and comparing these methods with some classical finite difference schemes. Several tests on real and virtual images will be presented.

3.16 Discrete surface processing and adaptive remeshing

Serena Morigi (University of Bologna, IT)


License  Creative Commons BY-NC-ND 3.0 Unported license
© Serena Morigi

We use various partial differential equation (PDE) models to efficiently solve several surface processing, including reconstruction, smoothing, and remeshing.

In particular, we propose a new adaptive remeshing strategy for the regularization of arbitrary topology triangulated surface meshes. Unlike existing sophisticated parametric remeshing techniques, our explicit method redistributes the vertices on the surface by keeping all edges on element stars approximately of the same size, and areas proportional to the surface features. At this aim we solve a two-step PDE model using discrete differential geometry operators suitably weighted to preserve surface curvatures and to obtain a good mesh quality, that is well-shaped triangles. Several examples demonstrate that the proposed approach is simple, efficient and gives very desirable results especially for surface models having sharp creases and corners.

3.17 A truly Unsupervised, Non-Parametric Clustering Method

Pablo Muse (Universidad de la Republica – Montevideo, UY)

License  Creative Commons BY-NC-ND 3.0 Unported license
© Pablo Muse

Joint work of Muse, Pablo; Tepper, Mariano; Almansa, Andrés

Human perception is extremely adapted to group similar visual objects. The Gestalt school studied the perceptual organization and identified a set of rules that govern human perception. One of the earlier and most powerful qualities, or gestalts, is proximity, which states that spatial or temporal proximity of elements may induce to perceive them as a single group. From an algorithmic point of view, the main problem with the gestalt rules is their qualitative nature. Our goal is to design a clustering method that can be considered a quantitative assessment of the proximity gestalt. We show that this can be achieved by analyzing the inter-point distances of the Minimum Spanning Tree, a structure that is closely related to human perception. We present a method that relies on the sole characterization of non-clustered data, thus being capable of detecting non-clustered data as such, and to detect clusters of arbitrary shape.


The method is fully unsupervised in the sense that the user input only relates to the nature of the problem to be treated, and not the clustering algorithm itself. Even the number

of clusters does not need to be previously chosen.

Strictly speaking the method involves one single parameter that controls the degree of reliability of the detected clusters. However, the algorithm can be considered parameter-free, as the result is not sensitive to its value.

3.18 Image-based 3D Modeling via Cheeger Sets

Martin Oswald (TU München, DE)


License  Creative Commons BY-NC-ND 3.0 Unported license
© Martin Oswald

Joint work of Oswald, Martin; Toeppe, Eno; Cremers, Daniel; Rother, Carsten

In this talk, we present a novel variational formulation for generating 3D models of objects from a single view. Based on a few user scribbles in an image, the algorithm automatically extracts the object silhouette and subsequently determines a 3D volume by minimizing the weighted surface area for a fixed user-specified volume. The respective energy can be efficiently minimized by means of convex relaxation techniques, leading to visually pleasing smooth surfaces within a matter of seconds. In contrast to existing techniques for single-view reconstruction, the proposed method is based on an implicit surface representation and a transparent optimality criterion, assuring high-quality 3D models of arbitrary topology with a minimum of user input.

3.19 Nested Sphere Statistics of Skeletal Models

Stephen M. Pizer (University of North Carolina – Chapel Hill, US)





License  Creative Commons BY-NC-ND 3.0 Unported license
© Stephen M. Pizer

We seek a form of object model that exactly and completely captures the interior of most non-branching anatomic models and simultaneously is well suited for probabilistic analysis on populations of such objects. We show that certain nonmedial, skeletal models satisfy these requirements. These models will first be mathematically defined in continuous 3-space, and then discrete representations will be derived. We will describe means of fitting skeletal models into manual or automatic segmentations of objects in a way stable enough to support statistical analysis, and we will specify means of modifying these fits to provide good correspondences of spoke vectors across a training population of objects. Understanding will be developed that these discrete skeletal models live in an abstract space made of a Cartesian product of a Euclidean space and a collection of spherical spaces. Based on this understanding and the way objects change under various rigid and nonrigid transformations, a method analogous to principal component analysis called composite principal nested spheres will be seen to apply to learning an efficient collection of modes of object variation about a Fréchet mean object.

The methods will be illustrated by application to a few anatomic objects.

3.20 Group-valued regularization for motion segmentation of dynamic non-rigid shapes





Guy Rosman (Technion – Haifa, IL)

License     Creative Commons BY-NC-ND 3.0 Unported license
© Guy Rosman

Understanding of articulated shape motion plays an important role in many applications in the mechanical engineering, movie industry, graphics, and vision communities. In this paper, we study motion-based segmentation of articulated 3D shapes into rigid parts. We pose the problem as finding a group-valued map between the shapes describing the motion, forcing it to favor piecewise rigid motions. Our computation follows the spirit of the Ambrosio-Tortorelli scheme for Mumford-Shah segmentation, with a diffusion component suited for the group nature of the motion model. Experimental results demonstrate the effectiveness of the proposed method in non-rigid motion segmentation.

3.21 Variational Models in Shape Space and Links to Continuum Mechanics





Martin Rumpf (Universität Bonn, DE)

License     Creative Commons BY-NC-ND 3.0 Unported license
© Martin Rumpf

The analysis of shapes as elements in a frequently infinite-dimensional space of shapes has attracted increasing attention over the last decade. The aim of this talk is to adopt a primarily physical perspective on the space of shapes and to relate this to the prevailing geometric perspective. Indeed, we consider shapes given as boundary contours of volumetric objects, which consist either of an elastic solid or a viscous fluid. In the first case, shapes are transformed via elastic deformations, where the associated elastic energy only depends on the final state of the deformation and not on the path along which the deformation is generated. The minimal elastic energy required to deform an object into another one can be considered as a dissimilarity measure between the corresponding shapes. We apply this approach for shape averaging and shape statistics. In the second case, shapes are transformed into each other via viscous transport of fluid material, and the flow naturally generates a connecting path in the space of shapes. The viscous dissipation rate, the rate at which energy is converted into heat due to friction, can be defined as a metric on an associated Riemannian manifold.

3.22 3D Curve Skeleton Computation and Use for Discrete Shape Analysis

Gabriella Sanniti Di Baja (CNR – Naples, IT)


License     Creative Commons BY-NC-ND 3.0 Unported license
© Gabriella Sanniti Di Baja

A discrete 3D curve skeletonization algorithm is described, based on iterated voxel removal guided by the distance transform of the object and on anchor point selection. The anchor

point selection criterion is adequate to originate a curve skeleton reflecting object's shape sufficiently well. Then, the use of the curve skeleton for object decomposition in the framework of the structural approach to shape analysis is discussed. A suitable partition of the skeleton is presented that originates object decomposition in accordance with human intuition.

3.23 Incremental Level Set Tracking


Nir Sochen (Tel Aviv University, IL)

License  Creative Commons BY-NC-ND 3.0 Unported license
© Nir Sochen

We consider the problem of contour tracking in the level set framework. Level set methods rely on low level image features, and very mild assumptions on the shape of the object to be tracked. To improve their robustness to noise and occlusion, one might consider adding shape priors that give additional weight to contours that are more likely than others. This works well in practice, but assumes that the class of object to be tracked is known in advance so that the proper shape prior is learned. In this work we propose to learn the shape priors on the fly. That is, during tracking we learn an eigenspace of the shape contour and use it to detect and handle occlusions and noise. Experiments on a number of sequences reveal the advantages of our method.

3.24 Global Minimization for Continuous Multiphase Partitioning Problems Using a Dual Approach and graph cuts algorithms

Xue-Cheng Tai (University of Bergen, NO)


License  Creative Commons BY-NC-ND 3.0 Unported license
© Xue-Cheng Tai

Joint work of Tai, Xue-Cheng; Baue, Egil; Yuan, Jing

This talk is devoted to the optimization problem of continuous multi- partitioning, or multi-labeling, which is based on a convex relaxation of the continuous Potts model. In contrast to previous efforts, which are trying to tackle the optimal labeling problem in a direct manner, we first propose a novel dual model and then build up a corresponding duality-based approach. By analyzing the dual formulation, we can show that the relaxation is often exact, i.e. the optimal solution is also globally optimal to the nonconvex Potts model. In order to deal with the nonsmooth dual problem, we suggest a smoothing method based on the log-sum exponential function and also indicate that such smoothing approach gives rise to the novel smoothed primal-dual model and suggests labelings with maximum entropy. Such smoothing method for the dual model produces an expectation maximization algorithm for the multi-labeling problem. Numerical experiments show competitive performance in terms of quality and efficiency compared to several state of the art methods for the Pott's model. In the end, we will also present several recent algorithms for computing global minimizers based on graph cut algorithms and augmented Lagrangian approaches.

3.25 Non-Local Ambrosio-Tortorelli and 3-Partite Skeletons

Sibel Tari (Middle East Technical University – Ankara, TR)


License  Creative Commons BY-NC-ND 3.0 Unported license
© Sibel Tari

The phase field of Ambrosio-Tortorelli, which results from a competition between phase separation and local neighborhood interaction, has proven to be an indispensable tool in variational formulations of shape analysis that jointly involve region and boundary terms. When the local neighborhood interaction is emphasized over the phase separation, a smooth distance function, which codes local symmetries hence skeletons, is obtained.

In this talk, I will introduce a non-local modification to the otherwise local interaction term by adding the L^2 norm of the expectation, which suggests further regularization of the smooth distance function by a zero-sum function. After discussing the modified field, I will show alternative skeleton constructions leading to 3-partite forms separately coding bodies, appendages, and boundary texture.

3.26 Integrated DEM Construction and Calibration of Hyperspectral Imagery: A Remote Sensing Perspective

Christian Woehler (TU Dortmund, DE)

License  Creative Commons BY-NC-ND 3.0 Unported license
© Christian Woehler

In this study we present an approach to the image-based construction of planetary digital elevation maps (DEM) by fusion of hyperspectral imagery with depth data obtained by laser altimetry measurements. Photometric methods like photoclinometry, shape from shading, or photometric stereo yield dense surface normal fields with a lateral resolution coming close to that of the images themselves, but the inferred relative depth data tend to be strongly biased on large scales, while laser altimetry provides absolute depth data with the drawback of high-frequency noise and limited lateral resolution. Our DEM construction algorithm applies an iterative scheme to combine absolute depth and gradient data. It initially operates on coarsely downsampled absolute depth data and then successively increases the resolution for each iteration step. During each step, the depth data and the image radiances on the respective spatial scale are used to estimate the non-uniform surface albedo, which then yields an updated DEM. The resulting DEM reveals fine surface details as its lateral resolution is comparable to that of the original images. We describe the surface reflectance by the physically motivated Hapke model, which is an analytical approximation to the radiative transfer equation.


Since for DEM construction we employ 85-band hyperspectral image data of 140 m lateral resolution provided by the Chandrayaan-1 Moon Mineralogy Mapper (M3) instrument, we are able to obtain for each hyperspectral pixel a reflectance spectrum normalised to a standard configuration of 30° incidence angle, 0° emission angle, and 30° phase angle, where in contrast to standard hyperspectral calibration approaches the detailed topography is taken into account. We point out the effect of the wavelength-dependent Hapke model based calibration on important features extracted from the normalised spectra (wavelength, relative depth, and width of the characteristic lunar iron absorption trough at 1000 nm; spectral ratio between 2817 and 2657 nm indicating water and/or hydroxyl ions).

However, a detailed inspection of the spectral features reveals pronounced residual dependences on topography-induced illumination variations, which are probably due to imperfections of the Hapke model of the order of one percent or less. Hence, we propose an empirical approach to learn these residual deviations from compositionally homogeneous surface regions displaying a large variety of slopes, such as the inner walls of small craters, based on the high-resolution DEM data. We show that our empirical correction strongly reduces topography-induced illumination effects on the reflectance spectra. As an ultimate refinement step, the corrected reflectances can be used to determine corrected albedo maps and DEM data.

Finally, we discuss open issues, specifically the physical modelling of topography-induced effects (e.g. by introducing a wavelength-dependent single-particle angular scattering function), and the fundamental problem of distinguishing between effects related to illumination vs. surface temperature at infrared wavelengths.

3.27 Stochastic Diffeomorphic Evolution and Tracking

Laurent Younes (Johns Hopkins University, US)

License  Creative Commons BY-NC-ND 3.0 Unported license
© Laurent Younes

We will describe how shape evolution can be controlled by “diffeomorphic finite elements”, or diffeons. In such an approach, diffeons, which are shape dependent vector fields, are combined linearly and integrated in a dynamical system to generate diffeomorphic motion. As such, they can be used in segmentation problems, where they lead (combined with gradient descent) to diffeomorphic active contours. They can also be generated as a stochastic process, leading to random shape evolution. We will discuss the definition of such random processes, and focus in particular on "Eulerian" definitions of the process (depending on the current shape location), in contrast with "Lagrangian" definitions that relate to fixed coordinate systems or parametrizations. We will show examples of shape generated by such processes and discuss preliminary applications in diffeomorphic shape tracking.

3.28 Shape Analysis for 3D Point Cloud.

Hong-Kai Zhao (University of California – Irvine, US)

License  Creative Commons BY-NC-ND 3.0 Unported license
© Hong-Kai Zhao


In the first part I will present an efficient algorithm for computing the Euclidean skeleton of an object directly from a point cloud representation on an underlying grid. The key point of this algorithm is to identify those grid points that are (approximately) on the skeleton using the closest point information of a grid point and its neighbors. The three main ingredients of the algorithm are: (1) computing closest point information efficiently on a grid, (2) identifying possible skeletal points based on the number of closest points of a grid point and its neighbors with smaller distances, (3) applying a distance ordered homotopic thinning process to remove the non-skeletal points while preserving the end points or the edge points of the skeleton.

In the second part I will present our recent work on present implicit surface reconstruction algorithms for point clouds. We view the implicit surface reconstruction as a three dimensional

binary image segmentation problem that segments the computational domain into an interior region and an exterior region while the boundary between these two regions fits the data points properly. The key points with using an image segmentation formulation are: (1) an edge indicator function that gives a sharp indicator of the surface location, and (2) an initial image function that provides a good initial guess of the interior and exterior regions. In this work we propose novel ways to build both functions directly from the point cloud data. We then adopt recent convexified image segmentation models and fast computational algorithms to achieve efficient and robust implicit surface reconstruction for point clouds.

3.29 Distance Images and Intermediate-Level Vision

Steven W. Zucker (Yale University, US)


License  Creative Commons BY-NC-ND 3.0 Unported license
© Steven W. Zucker

The early stages of computer vision are dominated by image patches or features derived from them; high-level vision is dominated by shape representation and recognition. However there is almost no work between these two levels, which creates a problem when trying to recognize complex categories such as “airports” for which early feature clusters are ineffective. We argue that an intermediate-level representation is necessary and that it should incorporate certain high-level notions of distance and geometric arrangement into a form derivable from images.

We propose an algorithm based on a reaction-diffusion equation that meets these criteria; we prove that it reveals (global) aspects of the distance map locally; and illustrate its performance on airport and other imagery, including visual illusions. Finally, we observe that the feedforward and feedback pathways that define the intermediate-levels of biological vision systems could benefit directly from such models, and we sketch one plausible path for implementing them via local field potentials.

3.30 Orientation and Anisotropy of Multicomponent Shapes

Jovisa Zunic (University of Exeter, GB)

License  Creative Commons BY-NC-ND 3.0 Unported license
© Jovisa Zunic

Orientation and anisotropy of multicomponent shapes will be considered. There are many situations where a number of single objects are better considered as components of a multicomponent shape (e.g. a fish shoal), but also there are situations where a single object is better segmented into natural components and considered as a multicomponent shape (decomposition of cellular materials onto the corresponding cells). These problems have not been considered previously, even though both orientation and anisotropy problems of single component shapes have been considered earlier.

Participants

- Fethallah Benmansour
EPFL - Lausanne, CH
- Michael Breuß
Univ. des Saarlandes, DE
- Alex M. Bronstein
Tel Aviv University, IL
- Michael M. Bronstein
Universität Lugano, CH
- Thomas Brox
Universität Freiburg, DE
- Alfred M. Bruckstein
Technion – Haifa, IL
- Elisabetta Carlini
University of Rome
“Sapienza”, IT
- Leila De Floriani
University of Genova, IT
- Stephan Didas
Fraunhofer ITWM –
Kaiserslautern, DE
- Anastasia Dubrovina
Technion – Haifa, IL
- Maurizio Falcone
University of Rome
“Sapienza” IT
- P. Thomas Fletcher
University of Utah, US
- Silvano Galliani
Univ. des Saarlandes, DE
- Hans-Christian Hege
ZIB – Berlin, DE
- Petros Maragos
National TU - Athens, GR
- Roberto Mecca
University of Rome
“Sapienza” IT
- Serena Morigi
University of Bologna, IT
- Pablo Muse
Universidad de la Republica –
Montevideo, UY
- Martin Oswald
TU München, DE
- Stephen M. Pizer
University of North Carolina
– Chapel Hill, US
- Guy Rosman
Technion - Haifa, IL
- Martin Rumpf
Universität Bonn, DE
- Gabriella Sanniti Di Baja
CNR – Naples, IT
- Nir Sochen
Tel Aviv University, IL
- Xue-Cheng Tai
University of Bergen, NO
- Sibel Tari
Middle East Technical
University – Ankara, TR
- Christian Wöhler
TU Dortmund, DE
- Laurent Younes
Johns Hopkins University, US
- Hong-Kai Zhao
University of California –
Irvine, US
- Steven W. Zucker
Yale University, US
- Jovisa Zunic
University of Exeter, GB



Formal Methods in Molecular Biology

Edited by

Rainer Breitling¹, Adelinde M. Uhrmacher², Frank J. Bruggeman³,
and Corrado Priami⁴

1 University of Glasgow, GB, rainer.breitling@glasgow.ac.uk

2 Universität Rostock, DE, lin@informatik.uni-rostock.de

3 CWI – Amsterdam, NL

4 Microsoft Research – University Trento, IT

Abstract

This report documents the program and the outcomes of the Seminar 11151 ‘Formal Methods in Molecular Biology’ that took place in Dagstuhl, Germany, on 10.–15. April, 2011. The most recent advances in Systems Biology were discussed, as well as and the contribution of computational formalisms to the modeling of biological systems, with the focus on stochasticity. About 30 talks were given. The participants formed 5 teams that worked on selected case studies. Two teams were awarded prizes, for their efforts in analyzing and further elucidating published biological models.

Seminar 10.–15. April, 2011 – <http://www.dagstuhl.de/11151>

1998 ACM Subject Classification J.3 Biology and Genetics.

Keywords and phrases bioinformatics, systems biology, formal modeling, computational biology, stochastic, model, simulation, checking, verification, abstraction, petri nets, process algebra

Digital Object Identifier 10.4230/DagRep.1.4.41

Edited in cooperation with Elzbieta Krepska

1 Executive Summary

Rainer Breitling

Adelinde M. Uhrmacher

License  Creative Commons BY-NC-ND 3.0 Unported license
© Rainer Breitling, Adelinde M. Uhrmacher

The second Dagstuhl Seminar on Formal Methods in Molecular Biology took place from 10–15 April, 2011. 35 participants from 8 countries gathered to discuss the most recent advances in Systems Biology and the contribution of computational formalisms to the successful modeling of biological systems. Major recurrent themes were the description of stochastic phenomena in biology, the modeling of spatial aspects of cellular behavior, and the robustness of cellular switches in the face of molecular noise and uncertainty of parameter inference. The computational modeling approaches applied to these challenges were particularly diverse, ranging from differential equation-based models to various flavors of rule-based languages, Petri Nets and process algebras.

A central component of the seminar was the Second International Biomodeling Competition. Teams formed during the first day and worked on biological case studies using a variety of modeling formalisms and analysis methods; the results were presented on Thursday afternoon and the winner determined by a joint vote of the audience.



Except where otherwise noted, content of this report is licensed under a Creative Commons BY-NC-ND 3.0 Unported license

Formal Methods in Molecular Biology, *Dagstuhl Reports*, Vol. 1, Issue 4, pp. 41–64

Editors: Rainer Breitling, Adelinde M. Uhrmacher, Frank J. Bruggeman, and Corrado Priami



Dagstuhl Reports

Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

The 1st prize went to the team of Kirill Batmanov, Antje Beyer, Matthias Jeschke and Carsten Maus, for their work on ‘Synchronization of cell populations’.

The 2nd prize went to the team of Andrea Bracciali, Mostafa Herajy, Pietro Lió, Chris Myers, Brett Olivier, and Natal van Riel for their work on ‘A bistable gene switch’.

Special prizes were awarded to the team of Chiara Bodei, Luca Bortolussi, Davide Chiarugi, Maria Luisa Guerriero, Jane Hillston Ivan Mura, Alberto Policriti, and Alessandro Romanel (for ‘Critical Analysis’), the team of Mary Ann Blätke, Qian Gao, David Gilbert, Simon Hardy, Monika Heiner, Andrzej Kierzek, Fei Liu and Wolfgang Marwan (for ‘Innovative use of Petri Nets’), and the team of Maciej Dobrzyński, Mathias John, Céline Kuttler, Bartek Wilczyński and Verena Wolf (for ‘A pure stochastic approach’).

2 Table of Contents

Executive Summary

<i>Rainer Breitling, Adelinde M. Uhrmacher</i>	41
--	----

Overview of Talks


Spatial modeling of the community effect <i>Kirill Batmanov</i>	45
Inductive Analysis when State Enumeration Explodes <i>Giampaola Bella</i>	45
JAK/STAT-Pathway - A Case Study of the Modular Petri Net Modeling Concept <i>Mary Ann Blätke</i>	46
Towards a Taxonomy of Causality-Based Biological Properties <i>Chiara Bodei</i>	47
Hybrid approximation of stochastic process algebra models of biological systems <i>Luca Bortolussi</i>	48
Modelling HIV infection: A computational overview <i>Andrea Bracciulli</i>	49
Metabolomics Systems Biology <i>Rainer Breitling</i>	49
Single cell signaling and protein expression noise give rise to digital response on the population level <i>Maciej Dobrzyński</i>	50
Tav4SB: analysis of kinetic models of biological systems <i>Anna Gambin</i>	50
Stochastic model of the plant circadian clock - A Bio-PEPA case study <i>Maria Luisa Guerriero</i>	51
Analysis of the regulatory motif dynamic of signaling networks using Petri net-based dynamic graphs <i>Simon Hardy</i>	51
Generalized Hybrid Petri Nets <i>Mostafa Herajy</i>	52
Equivalence and Discretisation in Bio-PEPA <i>Jane Hillston</i>	52
Investigating the Switch-like Response in Yeast Pheromone Signaling <i>Matthias Jeschke</i>	52
Reaction Rules with Constraints <i>Mathias John</i>	53
Understanding gene function by analysis of large scale molecular interaction network model behaviour <i>Andrzej M. Kierzek</i>	53
Proving stabilization of large-scale biological systems <i>Elzbieta Krepska</i>	54

Modeling the community effect in development <i>Céline Kuttler</i>	54
Colored Petri nets for modeling and analyzing biological systems <i>Fei Liu</i>	55
Automatic Reconstruction of Extended Petri Nets from Experimental Data <i>Wolfgang Marwan</i>	56
Multi-Level Rule Schemas <i>Carsten Maus</i>	56
Elementary, or not? <i>Ivan Mura</i>	57
Utilizing Stochastic Model Checking to Determine the Robustness of Genetic Circuits <i>Chris J. Myers</i>	57
What does it cost to be flexible? A constraint based approach to modelling a micro-organism in a changeable environment <i>Brett Olivier</i>	58
Rule-based modeling and application to biomolecular networks <i>Alessandro Romanel</i>	58
Abstractions in Spatial Simulation in Cell Biology <i>Adelinde M. Uhrmacher</i>	59
Combining Bayesian and Frequentist parameter inference methods in systems biology <i>Natal van Riel</i>	59
Cell lineage dynamics analysis with individual regulatory networks <i>Bartek Wilczyński</i>	59
Inference and Stability of Stochastic Models in Systems Biology <i>Verena Wolf</i>	60
Working Groups	
Team 1: DICTYPAT <i>Mary Ann Blätke, Qian Gao, David Gilbert, Simon Hardy, Monika Heiner, Andrzej Kierzek, Fei Liu, Wolfgang Marwan</i>	61
Team 2: A bistable gene switch <i>Andrea Bracciali, Mostafa Herajy, Pietro Lio, Chris J. Myers, Brett Olivier, Natal van Riel</i>	61
Team 4: Simulating coupled gene expression oscillations of cell populations during somitogenesis using delays <i>Antje Beyer, Kirill Batmanov, Matthias Jeschke, Carsten Maus</i>	62
Team 5: Counter-intuitive stochastic behavior of simple gene circuits with negative feedback <i>Ivan Mura, Davide Chiarugi, Maria Luisa Guerriero, Chiara Bodei, Luca Bortolussi, Alessandro Romanel, Jane Hillston, Alberto Policriti</i>	63
Participants	64

3 Overview of Talks

3.1 Spatial modeling of the community effect

Kirill Batmanov (*Université de Lille I, FR*)

License  Creative Commons BY-NC-ND 3.0 Unported license
© Kirill Batmanov

We investigated some spatial aspects of a phenomenon called community effect, first described in [1]. The community effect is said to be present in a system if it shows some behavior only when the number of cells in the system is large enough. To study the effects of diffusion on the community effect, we added one-dimensional space constraints on the diffusion to the model of the community effect described in [2]. We found the conditions which are necessary for the community effect in the new spatial model.


To study whether it is possible for the community effect to restrict itself in space through self-regulation, we built a simplified model based on Turing reaction-diffusion theory [3]. Using the gene network for the community effect in sea urchin described in [4] combined with the conditions for formation of Turing patterns [5], we showed by stochastic simulations that such a system indeed can become self-regulated and restrict its area of activation in space.

References

- 1 Gurdon JB, Tiller E, Roberts J, Kato K. *A community effect in muscle development*. *Current biology*, 3(1):1-11, 1993.
- 2 Saka Y, Lhoussaine C, Kuttler C, Ullner E. *Theoretical basis of the community effect in development*. To appear in *BMC Systems Biology*.
- 3 Turing AM. *The chemical basis of morphogenesis*. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 237(641):37, 1952.
- 4 Bolouri H, Davidson EH. *The gene regulatory network basis of the "community effect", and analysis of a sea urchin embryo example*. *Developmental biology*, 340(2):170-8, 2010.
- 5 Kondo S, Miura T. *Reaction-Diffusion Model as a Framework for Understanding Biological Pattern Formation*. *Science*, 329(5999):1616-1620, 2010.

3.2 Inductive Analysis when State Enumeration Explodes

Giampaolo Bella (*Universita di Catania, IT*)

License  Creative Commons BY-NC-ND 3.0 Unported license
© Giampaola Bella

Current computer science research in the domain of biology has helped establishing a suitable formalisation of biological networks as molecular-scale autonomous programmable computers that operate synchronous and asynchronous ‘logical’ control of biological processes. Each biological process can then be represented as a set of states amenable to model checking.

The computational paradigm underlying model checking is imperative in the sense that each state of the target system is modelled explicitly as an abstract state, that is a set of variables and their current values (also termed an algebra). The reasoning technique is the enumeration of all the possible states to check whether each satisfies a stated property. Clearly, the enumeration requires the system to be finite-state, but even so, very high number of states cannot be practically handled. Clever techniques exist to face this issue, and are led by symbolic model checking, which copes with up to 10^{20} states.

An effective reasoning technique for systems of unbounded size is mathematical induction. When a system is defined by induction, its properties can be assessed by the corresponding inductive proof principle. For example, induction is the natural way to define the set of the natural numbers, and inductive proofs can establish invariant properties of the naturals, such as the sum of the first n numbers. Remarkably, such proofs are independent from the size of n , and are also efficient because a functional computational paradigm is adopted: computation in fact proceeds by term rewriting, which uses the inductive hypothesis as an admissible rewriting rule. The efficiency can be practically realised by proving the mentioned $S_n = \frac{n(n+1)}{2}$ inductively as opposed to calculating S_n for a large n by enumerating the temporary states, which are sums in this case.


Theorem proving tools, also said proof assistants, offer great support to inductive reasoning. The properties of the genetic toggle switch have been studied inductively [1,2]. A theorem establishes stability, that both genes cannot prevail (in terms of concentration levels) at the same time. Attempting to prove that either gene cannot prevail over the other fails, hence enforcing the opposite, desirable property. This sort of meta-proving, together with the stability theorem, provide formal guarantee that the toggle is bistable. The potential to tackle much more complex biological systems, where the combinatorics becomes hardly manageable, appears huge and yet unexplored.

References

- 1 Bella, G., Liò, P.: *Formal analysis of the genetic toggle*. In Degano, P., Gorrieri, R., eds.: Proc. of the 9th Conference on Computational Methods in Systems Biology (CMSB'09). LNBI 5688, Springer (2009), 96–110.
- 2 Bella, G., Liò, P.: *Analysing the microRNA-17-92/Myc/E2F/RB Compound Toggle Switch by Theorem Proving*. In Romano, P., eds.: Proc. of the 9th Workshop on Network Tools and Applications in Biology (Nettab'09). Liberodiscrivere (2009), 59-62.

3.3 JAK/STAT-Pathway - A Case Study of the Modular Petri Net Modeling Concept

Mary Ann Blätke (Universität Magdeburg – IBIO, DE)

License  Creative Commons BY-NC-ND 3.0 Unported license
© Mary Ann Blätke





Here, we present a computational model of the JAK-STAT pathway in IL6-signaling, which describes the molecular mechanisms in a high resolution. In addition, we introduce our modular modeling concept, which is based on the Petri net formalism and a modular approach. In our modeling concept, every protein, its interactions and intramolecular changes, are represented by an independent submodel, called module. Therefore, each module integrates the wide-spread information about a protein. A comprehensive model can be assembled from a set of modules of interacting proteins. Advantageously, our concept itself is not at all limited to the JAK-STAT pathway. We also propose a platform to organize approved curated modules in order to allow the generation of molecular networks. This platform might be useful in bridging the gaps between experimental bioscientists and theoretically oriented systems biologists.

References

- 1 Blätke, M.-A., Marwan W. Modular and Hierarchical Modeling Concept for Large Biological Petri Nets Applied to Nociception. German Workshop on Algorithms and Tools for Petri Nets. Cottbus, Germany, 2010.
- 2 Blätke, M.-A. Petri-Netz Modellierung mittels eines modularen and hierarchischen Ansatzes mit Anwendung auf nozizeptive Signalkomponenten. Otto von Guericke University Magdeburg. 2010 (Diploma thesis).
- 3 Blätke, M.-A. et al. Petri Net Modeling via a Modular and Hierarchical Approach Applied to Nociception. Int.Workshop on Biological Processes & Petri Nets (BioPPN), satellite event of Petri Nets 2010. Braga, Portugal, 2010.

3.4 Towards a Taxonomy of Causality-Based Biological Properties

Chiara Bodei (*University of Pisa, IT*)

License     Creative Commons BY-NC-ND 3.0 Unported license
© Chiara Bodei

Main reference C. Bodei, A. Bracciali, D. Chiarugi, R. Gori, “A Taxonomy of Causality-Based Biological Properties,” Proc. Third Workshop From Biology To Concurrency and back (FBTC), 2010, pp. 116–133, EPTCS 19.

URL <http://dx.doi.org/10.4204/EPTCS.19.8>

We formally characterize a set of causality-based properties of metabolic networks. This set of properties aims at making precise several notions on the production of metabolites, which are familiar in the biologists’ terminology. From a theoretical point of view, biochemical reactions are abstractly represented as causal implications and the produced metabolites as causal consequences of the implication representing the corresponding reaction. The fact that a reactant is produced is represented by means of the chain of reactions that have made it exist. Such representation abstracts away from quantities, stoichiometric and thermodynamic parameters and constitutes the basis for the characterization of our properties. Moreover, we propose an effective method for verifying our properties based on an abstract model of system dynamics. This consists of a new abstract semantics for the system seen as a concurrent network and expressed using the Chemical Ground Form calculus. We illustrate an application of this framework to a portion of a real metabolic pathway. Some ideas on how to apply our approach to other models is presented.


The talk is mainly based on the paper [1].

References

- 1 C. Bodei, A. Bracciali, D. Chiarugi, R. Gori, “A Taxonomy of Causality-Based Biological Properties,” Proc. Third Workshop From Biology To Concurrency and back (FBTC), 2010, pp. 116–133, EPTCS 19.

3.5 Hybrid approximation of stochastic process algebra models of biological systems

Luca Bortolussi (*Università di Trieste, IT*)

License  Creative Commons BY-NC-ND 3.0 Unported license
© Luca Bortolussi

Models in systems biology tend to cluster around two families of mathematical tools: differential equations and stochastic processes. Even though, physically speaking, stochastic models have firmer grounds, their computational analysis is much more costly than that of their differential counterpart. In any case, ODE-based descriptions of biological systems are often valuable and provide deep insights. Indeed, it is known that, limiting to mass action models, ODE's are an approximation of the (average of) stochastic models, and the differences between the two vanish in the thermodynamic limit (i.e. when populations and system's size go to infinity). Recently, there have been many attempts to mix these two techniques, at least as far as simulation of biological systems is concerned, resulting in several hybrid simulation algorithms. Hybrid dynamical systems have also been a hot topic in the last two decades, with much research work spanning across the boundary between computer science and engineering control. The best known model among hybrid dynamical systems are hybrid automata. Stochastic extensions of such concept are also receiving recently much attention. In both cases, most of the interest is in the development of automated reasoning tools rather than in simulation. It is widely recognized that Computational Systems Biology can highly benefit from modeling approaches embodying some stochastic ingredient. A very popular line along which such incorporation is realized, is based on the use of stochastic process algebras, which are proposed as front-end languages to (automatically) generate mathematical models, usually Continuous Time Markov Chains (CTMC). Recently, such process algebra based languages have continuous, while keeping the other discrete and stochastic. The basic idea is to find the best trade off between accuracy and computational efficiency (stochastic simulations are much more expensive than ODE simulation).

In this talk we present a programme which aims to increase even more the flexibility of stochastic process algebras by providing them with a very general semantics based on (stochastic) hybrid systems, encompassing CTMC and ODE as special cases. Such an approach is motivated not only by the gain in flexibility, but also by the goal of exploiting, in a systematic manner, automated reasoning tools to provide as much information as possible from a given model. Our stochastic process algebra of choice is stochastic Concurrent Constraint Programming (sCCP), an extension of CCP in the stochastic setting. In addition to the standard CTMC-based semantics, we have also provided sCCP with an ODE-based semantics and with an hybrid automata based semantics [2]. In particular, we will present a semantics based on Stochastic Hybrid Automata [1], thereby guaranteeing the possibility of parameterizing the degree of continuity introduced in the model. The result is a lattice of hybrid automata models, that increasingly remove discreteness in favor of continuity. The approach allows also a dynamic reconfiguration of such degree, in accordance to properties of the current state of the system. This allows the description in a formal setting of different hybrid simulation strategies, opening the way for their use in the context of process algebra modelling.


We will discuss some biological examples, mentioning the problems involved in identifying the 'right' degree of discreteness (the best compromise between accuracy and efficiency), some mathematical results that guarantee correctness of the approximation (in a limit sense) [3], and the potential of hybrid models for managing uncertainty and for describing multi-scale systems.

References

- 1 L. Bortolussi and A. Policriti, 2009. Hybrid Semantics of Stochastic Programs with Dynamic Reconfiguration. Proceedings of Second International Workshop on Computational Models for Cell Processes - EINDHOVEN. 3rd November 2009. EPTCS. Vol.6., pp.63-76.
- 2 L. Bortolussi and A. Policriti, 2010. Hybrid Dynamics of Stochastic Programs. Theoretical Computere Science. Vol.411/20. pp.2052-2077.
- 3 L. Bortolussi, 2010. Limit behavior of the hybrid approximation of Stochastic Process Algebras. Lecture Notes in Computer Science. Vol.6148/2010. pp.367-381.

3.6 Modelling HIV infection: A computational overview

Andrea Bracciali (University of Stirling, UK)

License  Creative Commons BY-NC-ND 3.0 Unported license
© Andrea Bracciali


We survey a modelling experiment regarding HIV infection, which relies on computationally inspired modelling and analysis techniques. The results presented are a case of methodological cross comparison (stochastic and deterministic) and a novel implementation of model checking in therapy validation. On the methodology side, this work represents a paradigm of integrating formal methods and mathematical models as a general framework to study HIV multiple strains during disease progression and the associated therapies. Results ranges from traditional ODE-based deterministic analysis, such as sensitivity analysis, to Gillespie-based stochastic simulations, to a novel application of stochastic model checking to quantitatively measure and compare the efficacy of (idealised models of) HIV-related therapies.

References

- 1 A. Sorathiya, P. Lió and A. Bracciali ‘An integrated modelling approach for R5-X4 mutation and HAART therapy assessment’. Swarm Intelligence. Springer. 4(4), pages 319–340, 2010. (DOI: 10.1007/s11721-010-0046-4).
- 2 A. Sorathiya, P. Lió and A. Bracciali ‘Formal reasoning on qualitative models of coinfection of HIV and Tuberculosis and HAART therapy’. BMC Bioinformatics. 11(1): Asia Pacific Bioinformatics Conference. 2010.
- 3 M. Aldinucci, A. Bracciali, P. Lió, A. Sorathiya, and M. Torquati, ‘StochKit-FF: Efficient Systems Biology on Multicore Architectures’. Euro-Par 2010 Parallel Processing Workshops HeteroPar, HPC, HiBB, CoreGrid, UCHPC, HPCF, PROPER, CCPI, and VHPC Workshops. Ischia, Italy, 2010. Lecture Notes in Computer Science 6586. Springer.

3.7 Metabolomics Systems Biology

Rainer Breitling (University of Glasgow, UK)


License  Creative Commons BY-NC-ND 3.0 Unported license
© Rainer Breitling

Metabolomics, the comprehensive study of small molecules in a biological system, has a privileged position in Systems Biology; many of the most prominent success stories of the field relate to the analysis of metabolic systems, and metabolic changes are a much closer reflection of physiologically relevant differences in cellular function than, e.g., changes in global gene expression profiles. This presentation discussed a variety of methodological challenges faced

in large-scale metabolomic analysis, focussing on the collection of computational tools that we have developed for the processing, exploration and interpretation of the large data sets created in metabolomic experiments. It concludes by presenting a case study of metabolic modelling, which integrates detailed kinetic information on individual enzymatic reactions and the often noisy and uncertain data provided by metabolomics.

3.8 Single cell signaling and protein expression noise give rise to digital response on the population level

Maciej Dobrzyński (University College - Dublin, IE)


License  Creative Commons BY-NC-ND 3.0 Unported license
© Maciej Dobrzyński

Genetically identical cells can respond differently to extracellular stimuli and may therefore follow different fates within cellular population. The appearance of multiple distinct gene expression states is typically attributed to feedback regulation in nonlinear signaling networks which leads to bistability. Here we present a different mechanism, which also results in heterogeneous response between cells in a population but does not rely on feedback topology.

Our measurements of extracellular signal-regulated kinase (ERK) response to epidermal growth factor (EGF) in single HEK 293 cells suggest that the population displays digital behavior where cells assume either ON or OFF states. Based on experiments augmented with stochastic models and computational analysis, we argue that this phenomenon results from the interplay of variability in the amount of network components and the nonlinearity of the network response. Our results may be applicable to a wider class of problems in which single cells need to discretize noisy extracellular cues to achieve a robust response.

3.9 Tav4SB: analysis of kinetic models of biological systems


Anna Gambin (University of Warsaw, PL)

License  Creative Commons BY-NC-ND 3.0 Unported license
© Anna Gambin

Taverna Workbench (Hull et al., 2006) eases integration of tools for life science research in experiments described as workflows. We provide new services that extend the functionality of Taverna Workbench for systems biology. These services allow to perform numerical ODE simulations or model-checking of SBML (Hucka et al., 2003) models, and they allow to visualize the results of model analysis. As an example of usage we constructed exemplary workflows. Our services are executed in the newly created grid environment, which integrates heterogeneous software such as Mathematica (Wolfram Research, Inc., 2008, Mathematica Edition: Version 7.0), PRISM (Hinton et al., 2006) and SBML ODE Solver. The enduser goal of the Taverna services for Systems Biology (Tav4SB) project is to abstract details of the technology infrastructure ‘in the cloud’ which supports provided services. User guide, examples and all resources are available from <http://bioputer.mimuw.edu.pl/tav4sb> Web page.

3.10 Stochastic model of the plant circadian clock - A Bio-PEPA case study

Maria Luisa Guerriero (University of Edinburgh, UK)

License  Creative Commons BY-NC-ND 3.0 Unported license
© Maria Luisa Guerriero

Main reference Akman, O.E.; Guerriero, M.L.; Loewe, L. and Troein, C., “Complementary approaches to understanding the plant circadian clock,” Proc. Third Workshop From Biology To Concurrency and back (FBTC), 2010, pp. 1–19, EPTCS 19.

URL <http://dx.doi.org/10.4204/EPTCS.19.1>


Circadian clocks are oscillatory genetic networks that help organisms adapt to the 24-hour day/night cycle. The clock of the green alga *Ostreococcus tauri* is the simplest plant clock discovered so far. Its many advantages as an experimental system facilitate the testing of computational predictions. We present a model of the *Ostreococcus* clock in the stochastic process algebra Bio-PEPA and exploit its mapping to different analysis techniques, such as ordinary differential equations, stochastic simulation algorithms and model-checking. The small number of molecules reported for this system tests the limits of the continuous approximation underlying differential equations. We investigate the difference between continuous-deterministic and discrete-stochastic approaches. Stochastic simulation and model-checking allow us to formulate new hypotheses on the system behaviour, such as the presence of self-sustained oscillations in single cells under constant light conditions. We investigate how to model the timing of dawn and dusk in the context of model-checking, which we use to compute how the probability distributions of key biochemical species change over time. These show that the relative variation in expression level is smallest at the time of peak expression, making peak time an optimal experimental phase marker. This is joint work with Ozgur Akman, Laurence Loewe and Carl Troein, published in [1].

References

- 1 Akman, O.E.; Guerriero, M.L.; Loewe, L. and Troein, C., “Complementary approaches to understanding the plant circadian clock,” Proc. Third Workshop From Biology To Concurrency and back (FBTC), 2010, pp. 1–19, EPTCS 19.

3.11 Analysis of the regulatory motif dynamic of signaling networks using Petri net-based dynamic graphs


Simon Hardy (Mount Sinai Medical School – New York, US)

License  Creative Commons BY-NC-ND 3.0 Unported license
© Simon Hardy

Numerous signaling pathways have been discovered in the mammalian cell and they form a highly connected network. It is generally accepted that this signaling apparatus not only transmits information from sources to targets, but also processes and transforms signals using biological control devices known as regulatory motifs. The simpler motifs are loops and bifans, but multiple motifs can be aggregated together to form more complex processing systems. As a result, cellular modules can exhibit high-level behaviors like bistability, oscillation, pulse generation and noise filtering. We have developed the dynamic graph using Petri net theory to analyze the emergent behavior of interacting regulatory motifs and we will present some examples.

3.12 Generalized Hybrid Petri Nets

Mostafa Herajy (BTU Cottbus, DE)

License  Creative Commons BY-NC-ND 3.0 Unported license
© Mostafa Herajy


Recently hybrid modelling and simulation of biochemical systems have attracted increasing interest. This is motivated by the need of simulating systems which integrate different sub-cellular models, and the fact that bio networks themselves are inherently stochastic, however stochastic simulation is time expensive.

Compared to other methods of biological modelling, Petri nets are characterized by their intuitive visual representation and executability of biological models.

Generalized Hybrid Petri Nets (GHPN) are recently introduced into Snoopy to combine both continuous Petri nets and generalized stochastic Petri nets. The modelling power of this class of Petri nets combines stochastic, discretely timed and continuous reactions into one model, which permits representing biological switches, in which continuous elements are turned on/off based on discrete elements. Moreover the defined model can be simulated using static or dynamic partitioning. The implementation of GHPN is freely available as part of Snoopy. In this talk, we provide a short presentation of the GHPN's elements as well as how it can be used to represent and simulate stiff biochemical networks.

3.13 Equivalence and Discretisation in Bio-PEPA

Jane Hillston (University of Edinburgh, UK)


License  Creative Commons BY-NC-ND 3.0 Unported license
© Jane Hillston

Bio-PEPA is a process algebra for modelling biological systems. An important aspect of Bio-PEPA is the ability it provides to discretise concentrations resulting in a smaller, more manageable state space. The discretisation is based on a step size which determines the size of each discrete level and also the maximum number of levels.

This talk considers equivalence relations for Bio-PEPA, particularly an equivalence addressing the relationship between two discretisations of the same Bio-PEPA model that differ only in the step size and hence the maximum number of levels. We use the idea of bisimulation from concurrency and process algebra. We present a novel behavioural semantic equivalence, compression bisimulation, that equates two discretisations of the same model.

3.14 Investigating the Switch-like Response in Yeast Pheromone Signaling

Matthias Jeschke (Heidelberg, DE)

License  Creative Commons BY-NC-ND 3.0 Unported license
© Matthias Jeschke

Coming originally from the field of stochastic simulation, the models used in previous work were artificially constructed to ease the performance evaluation of simulation algorithms in terms of execution speed and accuracy. But results from those studies tell only half the story.

That is why I have decided to focus on the application site of my work, i.e. the development and analysis of real-world models, e.g., the pheromone signaling pathway found in yeast cells. Still at the beginning, the talk will provide a brief overview of this pathway and provide some details about selected components, especially the scaffold protein Ste5 and its crucial yet not entirely unraveled role in tuning the pathway response.

3.15 Reaction Rules with Constraints





Mathias John (IRI – Villeneuve d’Ascq, FR)

License     Creative Commons BY-NC-ND 3.0 Unported license
© Mathias John

Chemical reaction rules are the natural formalism for the stochastic modeling of biochemical systems. However, they only provide limited expressiveness. In this talk, I will present the modeling language React(C) that allows to impose constraints on the occurrence of reactions. I will roughly describe how rules with constraints allow to describe complex protein-protein interaction and to consider spatial aspects. I will underline the expressiveness of React(C) by scatching an encoding from the stochastic pi-calculus to React(C). I will also report on the price of this expressiveness in terms of complexity of simulation algorithms.

3.16 Understanding gene function by analysis of large scale molecular interaction network model behaviour

Andrzej M. Kierzek (University of Surrey, UK)


License     Creative Commons BY-NC-ND 3.0 Unported license
© Andrzej M. Kierzek

Understanding of how gene expression determines phenotype of the living cell under particular environmental conditions is one of the paramount goals of basic science. Prediction of the effects of gene activity perturbation on the behaviour of the living system is prerequisite for personalised molecular medicine and rational engineering in synthetic biology. Decades of the research in molecular biology resulted in the vast corpus of scientific articles on individual interactions between the molecular components in the living cells. The literature knowledge can be used to reconstruct molecular interaction networks and perform computational analysis of the cellular behaviour, where the gene function is analysed in the context of the global molecular machinery of the cell. Due to the time scale separation between gene regulation and metabolism it is useful to study linear models of the flux distribution in the metabolic networks. The genome scale metabolic reaction networks can be explored by linear programming to identify genes essential for different metabolic activities. The gene-metabolic phenotype relationship determined in this way can be further exploited to analyse transcriptomics data. Our recent publication (Bonde et al. PLoS Computational Biology, in press) shows applications to understanding of metabolic vulnerabilities of Mycobacterium tuberculosis. Despite the success of the genome scale metabolic network analysis, these models allow investigation of only about 1000 genes in the cell. Majority of the genes participate in processes other than metabolism that cannot be modelled in quasi-steady state framework. Recently, it has been demonstrated that useful approximation of large scale systems dynamics can be obtained by sampling token game trajectories in the Petri-net

representation of the molecular interaction networks. I will show preliminary results on how this approach can be integrated with flux balance analysis of metabolism towards modelling of genome scale molecular interaction networks capable of predicting function of majority of the genes in the cell.

3.17 Proving stabilization of large-scale biological systems

Elzbieta Krepska (Vrije Universiteit – Amsterdam, NL)

License  Creative Commons BY-NC-ND 3.0 Unported license
© Elzbieta Krepska


Main reference B. Cook, J. Fisher, E. Krepska, N. Piterman, “Proving stabilization of biological systems,” Proc. 12th Verification, Model Checking and Abstract Interpretation Conference (VMCAI’11), LNCS vol. 6538, pp. 134–149, 2011.

URL http://dx.doi.org/10.1007/978-3-642-18275-4_11

We describe an efficient procedure for proving stabilization of biological systems modeled as qualitative networks or genetic regulatory networks. For scalability, our procedure uses modular proof techniques, where state-space exploration is applied only locally to small pieces of the system rather than the entire system as a whole. Our procedure exploits the observation that, in practice, the form of modular proofs can be restricted to a very limited set. For completeness, our technique falls back on a non-compositional counterexample search. Using our new procedure, we have solved a number of challenging published examples, including: a 3-D model of the mammalian epidermis; a model of metabolic networks operating in type-2 diabetes; a model of fate determination of vulval precursor cells in the *C.elegans* worm; and a model of pair-rule regulation during segmentation in the *Drosophila* embryo. Our results show many orders of magnitude speedup in cases where previous stabilization proving techniques were known to succeed, and new results in cases where tools had previously failed.

3.18 Modeling the community effect in development

Céline Kuttler (Université de Lille I, FR)

License  Creative Commons BY-NC-ND 3.0 Unported license
© Céline Kuttler

Main reference Yasushi Saka, Cedric Lhoussaine, Céline Kuttler, Ekkehard Ullner and Marco Thiel, “Theoretical basis of the community effect in development,” *BMC Systems Biology* 2011, 5:54

URL <http://dx.doi.org/10.1186/1752-0509-5-54>

Background: Genetically identical cells often show significant variation in gene expression profile and behaviour even in the same physiological condition. Notably, embryonic cells destined to the same tissue maintain a uniform transcriptional regulatory state and form a homogeneous cell group. One mechanism to keep the homogeneity within embryonic tissues is the so-called community effect in animal development. The community effect is an interaction among a group of many nearby precursor cells, and is necessary for them to maintain tissue-specific gene expression and differentiate in a coordinated manner. Although it has been shown that the cell-cell communication by a diffusible factor plays a crucial role, it is not immediately obvious why a community effect needs many cells.


Results: In this work, we propose a model of the community effect in development, which consists in a linear gene cascade and cell-cell communication. We examined the properties of the model theoretically using a combination of stochastic and deterministic

modelling methods. We have derived the analytical formula for the threshold size of a cell population that is necessary for a community effect, which is in good agreement with stochastic simulation results.

Conclusion: Our theoretical analysis indicates that a simple model with a linear gene cascade and cell-cell communication is sufficient to reproduce the community effect in development. The model explains why a community needs many cells. It suggests that the community's long-term behaviour is independent of the initial induction level, although the initiation of a community effect requires a sufficient amount of inducing signal. The mechanism of the community effect revealed by our theoretical analysis is analogous to that of quorum sensing in bacteria. The community effect may underlie the size control in animal development and also the genesis of autosomal dominant diseases including tumorigenesis.

3.19 Colored Petri nets for modeling and analyzing biological systems

Fei Liu (BTU Cottbus, DE)

License  Creative Commons BY-NC-ND 3.0 Unported license
© Fei Liu

Petri nets are especially suitable for representing and modeling the concurrent, asynchronous, and dynamic behavior of biological systems. However, standard Petri nets do not scale, so they are restrained to modeling small systems. Therefore, we introduce the application of colored Petri nets in Systems Biology, which provide a possibility to model complex biological systems in a compact and scalable way.


In this talk, we will first give the motivation to use colored Petri nets for modeling and analyzing biological systems. Then we will describe some scenarios where colored Petri nets can contribute to Systems Biology, e.g. systems with repetitive components. Finally, we will introduce our colored Petri net modeling tool, Snoopy, by briefly describing its functionalities and features, especially for biological purpose.

References

- 1 F. Liu and M. Heiner. Colored petri nets to model and simulate biological systems. In Int. Workshop on Biological Processes and Petri Nets (BioPPN), satellite event of Petri Nets 2010, June 2010.
- 2 F. Liu and M. Heiner. Computation of enabled transition instances for colored petri nets. In Proc. 17th German Workshop on Algorithms and Tools for Petri Nets (AWPN 2010), volume 643 of CEUR Workshop Proceedings, pages 5–65. CEUR-WS.org, October 2010.
- 3 C. Rohr, W. Marwan, and M. Heiner. Snoopy - a unifying petri net framework to investigate biomolecular networks. *Bioinformatics*, 26(7):974–975, 2010.

3.20 Automatic Reconstruction of Extended Petri Nets from Experimental Data

Wolfgang Marwan (*Universität Magdeburg – IBIO, DE*)

License  Creative Commons BY-NC-ND 3.0 Unported license
© Wolfgang Marwan


Network inference methods reconstruct mathematical models of molecular or genetic networks directly from experimental data sets. A previously reported mathematical method [1] delivers all possible alternative minimal networks that are consistent with a given discrete time series data set. In combination with answer set programming the approach is computationally highly efficient [2]. The original method is exact as it does not involve any heuristic interaction with the reconstruction process, however it produces only simple place/transition Petri nets. We refined the previously published algorithm to consider catalysis and inhibition of the reactions that occur in the underlying network. The results of the reconstruction algorithm are encoded in the form of an extended Petri net involving control arcs. This allows the consideration of processes involving mass flow and/or regulatory interactions. As a non-trivial test case, the phosphate regulatory network of enterobacteria was reconstructed using in silico-generated time-series data sets on wild-type and in silico mutants. The new exact algorithm reconstructs extended Petri nets from time series data sets by finding all alternative minimal networks that are consistent with the data. It suggested biochemically meaningful alternative molecular mechanisms for certain reactions in the network. The algorithm is useful to combine data from wild-type and mutant cells and may potentially integrate physiological, biochemical, pharmacological, and genetic data in the form of a single model.

References

- 1 Marwan W, Wagler A, Weismantel R: A mathematical approach to solve the network reconstruction problem. *Mathematical Methods of Operations Research* 2008, 67(1):117–132.
- 2 Ostrowski, M., Schaub, T., Durzinsky, M., Marwan, W., Wagler, A.: Automatic network reconstruction using ASP. *Theory and Practice of Logic Programming* 2011, in press.

3.21 Multi-Level Rule Schemas


Carsten Maus (*Universität Rostock, DE*)

License  Creative Commons BY-NC-ND 3.0 Unported license
© Carsten Maus

Observations of biological systems behaviors are performed at different levels of organization, i.e. proteins, compartments, individual cells, and cell populations. In order to capture all relevant dynamics and to structure a system accordingly, one might need to combine different levels when modeling a biological system. Formal languages explicitly supporting multi-level modeling facilitate the description of such models. In this talk, I will present the idea of a formal rule- based language for modeling biological systems at multiple levels. Our approach allows to structure models in a nested hierarchical manner and to let the different levels influence each other via upward and downward causation, while its rule-based modeling metaphor close to the notation of chemical reactions makes it easy to understand the syntax and keeps models small and readable.

3.22 Elementary, or not?


Ivan Mura (Microsoft Research – University Trento, IT)

License  Creative Commons BY-NC-ND 3.0 Unported license
© Ivan Mura

This presentation discusses about the exactness of the Stochastic Simulation Algorithm proposed by Gillespie for the simulation of biochemical systems. It focuses on the accuracy issues that modelers may be confronted to when playing with the level of abstraction. In particular, we show that the core of Gillespie’s formulation, i.e. the negative exponential distribution of event reaction times, sets a precise amount of molecular noise in the models, which not necessarily corresponds to the one existing in the modeled system. The propagation of this noise may affect stochastic model results in hardly predictable ways.

3.23 Utilizing Stochastic Model Checking to Determine the Robustness of Genetic Circuits


Chris J. Myers (University of Utah – Salt Lake City, US)

License  Creative Commons BY-NC-ND 3.0 Unported license
© Chris J. Myers

When designing and analyzing genetic circuits, researchers are often interested in the probability of the system reaching a given state within a certain amount of time. Usually, this involves simulating the system to produce some time series data and analyzing this data to discern the state probabilities. However, as the complexity of models of genetic circuits grow, it becomes more difficult for researchers to reason about the different states by looking only at time series simulation results of the models. To address this problem, this talk describes how to use stochastic model checking, a method for determining the likelihood that certain events occur in a system, with continuous stochastic logic (CSL) properties to obtain similar results. This goal is accomplished by the introduction of a methodology for converting a genetic circuit model (GCM) into a labeled Petri net (LPN). The state space of the LPN is then computed, and the resulting continuous-time Markov chain (CTMC) is analyzed using transient Markov chain analysis to determine the likelihood that the circuit satisfies a given CSL property in a finite amount of time. This talk illustrates a use of this methodology to determine the likelihood of failure in a genetic toggle switch and other circuits, and it compares these results to stochastic simulation-based analysis of the same circuits. Our results show that this method results in a substantial speedup as compared with conventional simulation-based approaches.

3.24 What does it cost to be flexible? A constraint based approach to modelling a micro-organism in a changeable environment

Brett Olivier (Free University – Amsterdam, NL)

License  Creative Commons BY-NC-ND 3.0 Unported license
© Brett Olivier

Constraint based modelling is a widely used methodology used to analyse and study biological networks on both a small and whole organism (genome) scale. Typically these models are underdetermined and constraint based methods (e.g. linear, quadratic optimization) are used to optimise specific model properties. This is assumed to occur under a defined set of constraints (e.g. stoichiometric, metabolic) and bounds (e.g. thermodynamic, experimental and environmental) on the values that the solution fluxes can obtain.

Perhaps the most well known (and widely used) analysis method is Flux Balance Analysis (FBA) where for a model a target flux is maximised (typically a flux to biomass) where the other input/output fluxes have been bound to simulate a single set of defined environmental conditions. However, in the wild, such an organism may experience continuous changes in state that arise from sources either externally (e.g. a change in nutrient supply) or internally such as a mutation (deletion) in a particular gene which leads to a concomitant loss of (or large change in) cellular function.

In this presentation I will be discussing how we are attempting to extend established constraint based approaches to include micro-organisms living in a changeable environment. The question of what an organism can do in order to become more (or less flexible) in such an environment has necessitated the development of new theory, models, software tools and even a proposed standard for model exchange.

3.25 Rule-based modeling and application to biomolecular networks


Alessandro Romanel (ENS – Paris, FR)

License  Creative Commons BY-NC-ND 3.0 Unported license
© Alessandro Romanel

Modelers of molecular signaling networks must cope with the combinatorial explosion of protein states generated by post-translational modifications and complex formation. Rule-based models provide a powerful alternative to approaches that require an explicit enumeration of all possible molecular species of a system. Such models consist of formal rules stipulating the (partial) contexts for specific protein-protein interactions to occur. These contexts specify molecular patterns that are usually less detailed than molecular species. Yet, the execution of rule-based dynamics requires stochastic simulation, which can be very costly. We briefly introduce some recent results on a formal abstract interpretation-based method to convert a rule-based model into a reduced system of differential equations and highlight actual research directions.

3.26 Abstractions in Spatial Simulation in Cell Biology


Adelinde M. Uhrmacher (Universität Rostock, DE)

License  Creative Commons BY-NC-ND 3.0 Unported license
© Adelinde M. Uhrmacher

In modelling and simulation of cell biological processes, spatial homogeneity in the distribution of components is a common but not always valid assumption. Thus, more and more spatial simulation algorithms are developed. To keep the calculation costs at bay some trade accuracy for execution speed, others combine different algorithms. The talk discusses two approaches most recently developed at our laboratory and the abstractions they are based upon.

3.27 Combining Bayesian and Frequentist parameter inference methods in systems biology


Natal van Riel (TU Eindhoven, NL)

License  Creative Commons BY-NC-ND 3.0 Unported license
© Natal van Riel

Parameter estimation in systems biology models is in general an ill-posed inverse problem. Many different combinations of parameter values yield a model that can describe the data equally well. Non-identifiable model parameters hamper the development of predictive models. Strategies for parameterization that consider parameter identifiability are necessary. A strategy for model identification is proposed, based upon a combination of Monte Carlo Multiple Minimization, Profile Likelihood analysis and Monte Carlo Markov Chains. The approach can diagnose potential pitfalls regarding model parameterization. This is important information to have prior to a full Bayesian analysis. The analyses results mainly focus on model predictions rather than parameter values. The approach is applied to parameterize and analyze a model of the JAK-STAT signaling pathway.

3.28 Cell lineage dynamics analysis with individual regulatory networks

Bartek Wilczyński (University of Warsaw, PL)

License  Creative Commons BY-NC-ND 3.0 Unported license
© Bartek Wilczyński

Understanding the dynamics of proliferating cells in tissues is one of the key challenges for the modeling community. Currently, most widely used models assume large number of cells and use differential equations for modeling population sizes of different cell subtypes within tissue. Recently, a very interesting work of Lander et al. [1] showed what types of negative feedback are preferred for the stability of the regeneration of an epithelium consisting of cells at 3 different stages of specification. Their results are supported by experiments indicating that changing concentration of certain gene products affects proliferation and specification rates in cell cultures.

The question we are asking is whether the behavior predicted and observed results in study can be reproduced with a model working on the level of gene regulation. We know that the gene regulation is the process underlying cell specification, but current models of

gene regulatory networks usually focus on the autonomous system within one cell and are not applicable to cell population analysis. However, while typical Boolean network models are more useful in stable state analysis, stochastic modeling frameworks, such as Stochastic Logical Networks [2], allow for simulation of multiple regulatory networks with adjustable transition rates.


In this work, we propose a regulatory network capable of representing the cell lineage system described by Lander et al. [1] with all its key components: linear progression through 3 different stages with cell death and simple signaling. We show how this system can be simulated in the SLN framework leading to a Continuous Time Markov Chain (CTMC) model specified by very few parameters and able to reproduce typical trajectories of the original dynamical model.

References

- 1 A.D. Lander, K.K. Gokoffski, FY Wan, Q. Nie, and A.L. Calof. Cell lineages and the logic of proliferative control. *PLoS Biol*, 7(1):e1000015, 2009.
- 2 B. Wilczyński and J. Tiuryn. Regulatory network reconstruction using stochastic logical networks. In *Computational Methods in Systems Biology*, pages 142–154. Springer, 2006.

3.29 Inference and Stability of Stochastic Models in Systems Biology

Verena Wolf (Universität des Saarlandes, DE)

License  Creative Commons BY-NC-ND 3.0 Unported license
© Verena Wolf

Main reference A. Andreychenko, L. Mikeev, D. Spieler, and V. Wolf, “Parameter Identification for Markov Models of Biochemical Reactions,” Proc. 23rd International Conference on Computer Aided Verification (CAV’11), LNCS vol. 6806, pp. 83–98, 2011.

URL http://dx.doi.org/10.1007/978-3-642-22110-1_8

In this talk I discuss two common problems of stochastic models in Systems Biology. The first problem is the inference of stochastic reaction rate constants from noisy observations at certain arbitrarily spaced observation time intervals. I will present a numerical method for the estimation of the rate constants based on the maximum likelihood method. The second problem is the stability analysis of stochastic models and in particular the analysis of multistable models. I will explain why deterministic approximations are not useful for studying stability properties and suggest two methods to approximate the equilibrium distribution of the stochastic model. The first method is based on drift arguments while the second one relies on a partial fluid approximation.

4 Working Groups

4.1 Team 1: DICTYPAT

Mary Ann Blätke, Qian Gao, David Gilbert, Simon Hardy, Monika Heiner, Andrzej Kierzek, Fei Liu, Wolfgang Marwan

License © © ⊖ Creative Commons BY-NC-ND 3.0 Unported license
© Mary Ann Blätke, Qian Gao, David Gilbert, Simon Hardy, Monika Heiner, Andrzej Kierzek, Fei Liu, Wolfgang Marwan

This talk project is an exercise in modelling multiple spatial scales via the dynamics of cooperative biological entities, using as a test case pattern formation in *Dictyostelium discoideum*.

We developed a generic approach to support the systematic modelling of multiscale biological systems by the use of colour in Petri nets, which promises to be particularly helpful in investigating spatial aspects of the behaviour of biochemical systems, such as communication and behaviour at the intra and intercellular levels.

We represent space using a regular grid and developed a generic network pattern expressed as a coloured Petri net which can be easily configured for one, two or three dimensional grid scenarios of varying grid size. The pattern clearly separates intra-cellular and inter-cellular behaviour to allow for refinements of the pattern component which corresponds to specific cell behaviour.

We have validated our approach by performing some initial computational experiments. However, further experiments are required due to their computational expense which is beyond what can be done within the time limits of this workshop.

4.2 Team 2: A bistable gene switch

Andrea Bracciali, Mostafa Herajy, Pietro Lio, Chris J. Myers, Brett Olivier, Natal van Riel

License © © ⊖ Creative Commons BY-NC-ND 3.0 Unported license
© Andrea Bracciali, Mostafa Herajy, Pietro Lio, Chris J. Myers, Brett Olivier, Natal van Riel

Our talk considered the model a bistable gene switch proposed by Mehra et al. [1]. Utilizing our combined expertise, we attempted to model this system using a variety of methods including ODE's using Sundials CVode, Matlab, and PySCes, stochastic methods using iBioSim, generalized hybrid Petri nets using Snoopy, and stochastic process algebra using BioPEPA. Our initial attempts were hampered by the lack of a critical parameter, the basal rate of production for the activator. ODE simulation was used though to determine the effect of varying this parameter, and a reasonable value was determined. ODE simulation was further utilized to show the effect of the external concentration control signal as well as the dynamical behavior of the switch (noting, for instance that when the control signal is removed, the system switches back to the initial state). ODE analysis was utilized to analyze the effect of the growth rate on stability of the switch which was also found to be very critical. In addition to reproduction of the paper's results, we developed alternative models using Snoopy, Bio-PEPA, and iBioSim. Using iBioSim, we were able to show how automatic abstraction can reduce stochastic simulation time substantially (i.e., from 6 minutes to 0.6 seconds for 100 SSA runs). Using stochastic simulation of this abstracted model, we are able to show the probability of switching over time for different initial conditions and parameter values. We also considered an alternative model in which the activator does not need to


bind to the repressor, and we demonstrated that this model has a much sharper switching behavior. Finally, we constructed a population model for visualization purposes, showing how a multitude of individuals behave coherently with different assumptions on the concentration of the control signal. Overall, this effort not only reproduced previous results, but it also produced new analyses leading to new insights into this system. We hope to further this analysis in the future.

References

- 1 Mehra S, Charaniya S, Takano E, Hu W-S (2008) A Bistable Gene Switch for Antibiotic Biosynthesis: The Butyrolactone Regulon in *Streptomyces coelicolor*. PLoS ONE 3(7): e2724.

4.3 Team 4: Simulating coupled gene expression oscillations of cell populations during somitogenesis using delays

Antje Beyer, Kirill Batmanov, Matthias Jeschke, Carsten Maus

License  Creative Commons BY-NC-ND 3.0 Unported license
© Antje Beyer, Kirill Batmanov, Matthias Jeschke, Carsten Maus

Somitogenesis is the process of formation of somites, the segmented blocks of mesoderm which will eventually differentiate into other tissues.

We first reproduced the results reported in [1] using the open-source mathematics software SAGE and an additional package for solving delay differential equations (DDEs). To test whether intrinsic noise plays a significant role in maintaining sustained oscillations and the synchronization between cells, we simulated the model stochastically using two different delay representations: (i) a certain number of consecutive exponentially distributed events and (ii) fixed intervals. A stochastic simulation essentially exhibits the same behaviour as the system of DDEs, but additionally shows synchronization between coupled cells in an extended three-cell model.

A new rule-based multi-level language [2] was applied to integrate the non-elementary rate laws into a coupled multi-cellular model. First results showed a weak performance of the prototype simulator in case of larger cell populations. To study the behavior of many cells with local coupling, we implemented the delay stochastic simulation algorithm described in [3]. Simulations show that despite local differences in oscillations the coupling is able to synchronize the cells globally in 1- and 2-dimensional configurations.

A hybrid model was built by integrating the stochastic simulation algorithm into a physical model of cell movement based on previous work [4] using the Molecular Dynamics framework [5]. We were able to show how cell divisions cause a movement of the PSM. As cells dropped out of PSM to become somites, they were arrested at different gene levels resulting in a distinct pattern formation.

References

- 1 Lewis, J. *Autoinhibition with transcriptional delay: a simple mechanism for the zebrafish somitogenesis oscillator*. Cur. Biol. 13, pp. 1398–1408, 2003.
- 2 Maus, C. et al. *Towards Rule-Based Multi-Level Modeling*. Poster, ICSB'10, 2010.
- 3 Schlicht R, Winkler G. *A delay stochastic process with applications in molecular biology*. J. Math. Biol., 57(5), pp. 613–48, 2008.
- 4 Beyer, A. et al. *A Dynamic Physical Model of Cell Migration, Differentiation and Apoptosis in Caenorhabditis elegans*. Proc. of the 11th ICSB, Springer, 2011 (under review).

- 5 Alder, B.J., Wainwright, T.E. *Studies in Molecular Dynamics. I. General Method.* J. Chem. Phys., 31(2), p. 459, 1959.

4.4 Team 5: Counter-intuitive stochastic behavior of simple gene circuits with negative feedback

Ivan Mura, Davide Chiarugi, Maria Luisa Guerriero, Chiara Bodei, Luca Bortolussi, Alessandro Romanel, Jane Hillston, Alberto Policriti

License © ⓘ ⊖ Creative Commons BY-NC-ND 3.0 Unported license
© Ivan Mura, Davide Chiarugi, Maria Luisa Guerriero, Chiara Bodei, Luca Bortolussi, Alessandro Romanel, Jane Hillston, Alberto Policriti

Results: Gene expression is a fundamentally stochastic process with randomness in transcription and translation. Usually, stochastic modelling techniques used in gene expression are based on the Gillespie approach, where the role of parameters and the level of modelling abstraction are essential. Even in a simple model of gene regulation different combinations of parameter sets and abstraction levels can lead to completely different dynamics. Moreover, these combinations can have a crucial impact on the computational feasibility of the Gillespie approach. We have considered a simple model of gene regulation where a transcriptional repressor negatively regulates its own expression and we have investigated to what extent hybrid approaches and QSSA can be useful. In particular, the characteristic dynamics emerging from the randomness in transcription and translation can be captured also by hybrid approaches and we have shown that for different biological realistic parameter sets, different combinations of hybrid approaches and QSSA cannot only preserve the stochastic dynamics, but also speed up simulation times.

Conclusions: We have shown that both mRNA and protein degradation play a role in noise control and that, in general, there can be multiple control points in feedback loops.

In particular, no single parameter's variation is responsible for feedback loop intensity but a (linear? non-linear?) combination of them.

References

- 1 Tatiana T. Marquez-Lago & Jörg Stelling: Counter-intuitive stochastic behavior of simple gene circuits with negative feedback. *Biophys. J.* (2010) 98: 1742–1750.

Participants

- Kirill Batmanov
Université de Lille I, FR
- Giampaolo Bella
Università di Catania, IT
- Antje Beyer
University of Cambridge, UK
- Mary Ann Blätke
Univ. Magdeburg – IBIO, DE
- Chiara Bodei
University of Pisa, IT
- Luca Bortolussi
Università di Trieste, IT
- Andrea Bracciali
University of Stirling, UK
- Rainer Breitling
University of Glasgow, UK
- Davide Chiarugi
University of Siena, IT
- Maciej Dobrzyński
University College – Dublin, IE
- Anna Gambin
University of Warsaw, PL
- David Gilbert
Brunel University, UK
- Maria Luisa Guerriero
University of Edinburgh, UK
- Simon Hardy
Mount Sinai Medical School –
New York, US
- Monika Heiner
BTU Cottbus, DE
- Mostafa Herajy
BTU Cottbus, DE
- Jane Hillston
University of Edinburgh, UK
- Matthias Jeschke
Heidelberg, DE
- Mathias John
IRI - Villeneuve d'Ascq, FR
- Andrzej M. Kierzek
University of Surrey, UK
- Elzbieta Krepaska
Vrije Univ. – Amsterdam, NL
- Céline Kuttler
Université de Lille I, FR
- Pietro Lió
University of Cambridge, UK
- Fei Liu
BTU Cottbus, DE
- Wolfgang Marwan
Univ. Magdeburg – IBIO, DE
- Carsten Maus
Universität Rostock, DE
- Ivan Mura
Microsoft Research – University
Trento, IT
- Chris J. Myers
University of Utah - Salt Lake
City, US
- Brett Olivier
Free Univ. – Amsterdam, NL
- Alberto Policriti
Università di Udine, IT
- Alessandro Romanel
ENS – Paris, FR
- Adelinde M. Uhrmacher
Universität Rostock, DE
- Natal van Riel
TU Eindhoven, NL
- Bartek Wilczyński
University of Warsaw, PL
- Verena Wolf
Universität des Saarlandes, DE



Challenges in Document Mining

Edited by

Hamish Cunningham¹, Norbert Fuhr², and Benno Stein³

1 University of Sheffield, UK, gate.ac.uk/hamish

2 Universität Duisburg-Essen, Germany, norbert.fuhr@uni-due.de

3 Bauhaus-Universität Weimar, Germany, benno.stein@uni-weimar.de

Abstract

This report documents the programme and outcomes of the Dagstuhl Seminar 11171 *Challenges in Document Mining*. Our starting point was the observation that document mining techniques are often applied in an isolated manner, with the consequence that their potential is still to be fully realised. The goal of the seminar was to analyze this untapped potential. To this end researchers from the main areas of document mining were invited to present their views, to synthesise an understanding of where and how the latest disciplinary achievements can be combined, and to develop a more integrative view on the state of the art and the prospects for future progress.

Seminar 25.–29. May, 2011 – www.dagstuhl.de/11171

1998 ACM Subject Classification H.1.2 [Information Systems: User/Machine Systems]; H.3.1 [Information Systems: Content Analysis and Indexing]; H.3.3 [Information Systems: Information Search and Retrieval]; I.2.7 [Computing Methodologies: Learning]; I.2.7 [Computing Methodologies: Natural Language Processing]; I.5.3 [Computing Methodologies: Clustering]

Keywords and phrases Cluster analysis, HCI, Retrieval models, Social mining and search, Semi-supervised learning

Digital Object Identifier 10.4230/DagRep.1.4.65


Edited in cooperation with Melikka Khosh Niat

1 Executive Summary

Hamish Cunningham

Norbert Fuhr

Benno Stein

License  Creative Commons BY-NC-ND 3.0 Unported license
© Hamish Cunningham, Norbert Fuhr, Benno Stein

About Document Mining

Document mining is the process of deriving high-quality information from large collections of documents like news feeds, databases, or the Web. Document mining tasks include cluster analysis, classification, generation of taxonomies, information extraction, trend identification, sentiment analysis, and the like. Although some of these tasks have a long research history, it is clear that the potential of document mining is still to be fully realised.

Part of the problem is that relevant document mining techniques are often applied in an isolated manner, addressing – from a user perspective – only a part of a task. For example, an intelligent cluster analysis requires adequate document models (from information retrieval) that are combined with sensible merging algorithms (from unsupervised learning), complemented by an intuitive labelling (from information extraction, natural language processing).



Except where otherwise noted, content of this report is licensed under a Creative Commons BY-NC-ND 3.0 Unported license
Challenges in Document Mining, *Dagstuhl Reports*, Vol. 1, Issue 4, pp. 65–99
Editors: Hamish Cunningham, Oren Etzioni, Norbert Fuhr, and Benno Stein



DAGSTUHL REPORTS
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

The deficit that we observe may also be understood as a lack of application and user orientation in research. For example, given a result set clustering task, users expect:

1. as many clusters as they identify topics in the result set,
2. that the documents within each cluster are semantically similar to each other, and
3. that each cluster is labeled intuitively.

In order to achieve such a satisfying solution, the state-of-art of concepts and algorithms from information retrieval, unsupervised learning, information extraction, and natural language processing have to be combined in a user-focussed manner.

Goals of the Seminar

The general idea was to take an overview the state of the art in document mining research and to define a research agenda for further work. Since document mining tasks are not tackled by a single technology, we wanted to bring a sample of the leading teams together and look at the area from a multidisciplinary point of view. In particular, the seminar should focus on the following questions:

- What are the relevant document mining tasks? The expectations and the potential for document mining changed significantly over time. Influential in this connection is the discovery of the enormous contributions of users to the Web, among others in the form of blogs, comments and reviews, as highly valuable information source.
- What are the options and limitations of cluster analysis? A major deal of cluster analysis research has been spent to merging principles and algorithms; today, and especially in document mining, the focus is on tailored document models, user integration, topic identification and cluster labelling, on the combination with retrieval technology (e.g. as result set clustering). Especially non-topical classification tasks attracted interest in this connection, such as genre classification, sentiment analysis, or authorship grouping. Moreover, theoretical foundations of cluster analysis performance in document mining as well as commonly accepted optimality measures are open questions.
- What are the document mining challenges from a machine learning perspective? A crucial constraint is the lack of sufficient amounts of labelled data. This situation will become even more unbalanced in the future, and current research—to mention domain transfer learning and transductive learning—aim at the development of technology to exploit the huge amount of unlabelled data to improve supervised classification.
- How will NLP and IE affect the development of the field? The use of NLP and IE in document mining is a success factor of increasing importance for document mining. NLP contributes technology for document modelling, style quantification, document segmentation, topic identification, and various information extraction and semantic annotation tasks. In this regard authorship and writing style modelling is still coming of age; this area forms the heart for high-level document mining tasks such as plagiarism analysis, authorship attribution, and information quality assessment.
- Are new interaction paradigms on the rise? Interface design and visualization are very important for effective user access to the output of the document mining process. Moreover, interactive document mining approaches like e.g. scatter-gather clustering pose new challenges for both the interface and the backend.
- How to evaluate and compare the different research efforts? Evaluation is essential for developing any kind of data mining method. So far, mainly system-oriented evaluation approaches have been used, where the data mining output is compared to some “gold

Program of the Dagstuhl Seminar 11171, 25.-29. May, 2011

	Tuesday	Wednesday	Thursday	Friday
9:00 - 10:00	Welcome, Talks	Talks	Talks	Working group presentation
10:30 - 11:00	Coffee break			
11:00 - 12:30	Talks	Working group topics	Talks	Wrapup
12:30 - 14:00	Lunch			
14:00 - 15:30	Talks	Working groups	Social event: Excursion to Trier	
15:30 - 16:00	Coffee break			
16:00 - 18:00	Talks	Working groups		
18:00 - 19:00	Dinner			
19:00 - 22:00	Demos	Working groups	Open work	

standard”. There is a lack of user-oriented evaluations (e.g. observing users browsing a cluster hierarchy), that also take into account the tasks the users want to perform—e.g. using Borlund’s concept of simulated work tasks.

Seminar Organization

To stimulate debate and cross-pollination we scheduled a mixture of of talks, working groups and demos. Following Dagstuhl tradition, the talks were characterized by interactive discussions and provided a platform for presenting and discussing new ideas. The working group topics were arrived at by a brainstorming session. Due to Easter Monday we had only four days for our seminar and shifted parts of the program to the evening. The table shows the schedule of our seminar at a glance.

Selected Results

This week showed that there is a number of recurring themes that are addressed by different researchers:

1. The processing hierarchy: Classic methods in document mining deal with document clustering and classification, thus regarding documents as a whole (or an “atomic” unit). Recently, researchers have become interested in deeper analyses of texts, such as sentiment analysis and the extraction of entities and relations.
2. Unsupervised vs. supervised methods: The former can be applied easily, but often lead only to modest results. Supervised methods produce more valuable results, but require large training sets for generating high-quality output. However the two approaches are not real alternatives: there are various attempts for their combination, like e.g. using prior knowledge for improving clustering, or using unclassified data with clustering for classification.
3. Whereas most supervised methods are strongly domain-dependent, there are now attempts for developing more domain-independent or cross-domain methods that can be applied more universally.
4. User feedback and user interaction has become an important component of document mining: There are many approaches aiming at better visualizations of the mining results. More recently, visual analytics methods have become popular, which aim at supporting

the user during the mining process itself, thus incorporating the user's knowledge in the actual analysis (and not only during the training stage or for result presentation).

The week also showed (or confirmed) deficiencies in document mining and pointed to future research directions:

1. The multiplicity of retrieval questions, mining solutions, test corpora and evaluation measures (to mention only a few determinants) emerges naturally when satisfying individual information needs in our information-flooded society. However, this multiplicity hinders the comparison of solutions and hence the consequent improvement of the most promising technology. What is the ideal research infrastructure to exploit synergies?
2. That we suffer from an information overload is a commonplace. We ask: A single person has no chance to process a substantial part of the information at our disposal—but a machine can. How can we benefit from this fact?
3. Current retrieval and mining technology is text-centered. The question is if and how the respective machinery can be applied to complex objects and artificially generated data: Which elements of the state-of-the-art retrieval technology is of generic type, and, can we develop a retrieval theory for complex structures?

First answers and arguments related to these questions can be found in the working groups section of this report.

2 Table of Contents

Executive Summary

<i>Hamish Cunningham, Norbert Fuhr, Benno Stein</i>	65
---	----

Overview of Talks


Knowledge Discovery in the Web: Potential, Automation and Limits <i>Stefan M. Rüger</i>	71
Full Lifecycle Information Extraction and Multiparadigm Indexing with GATE <i>Hamish Cunningham</i>	71
Unsupervised and Semi-Supervised Approaches to Cross-Domain Sentiment Classification <i>John A. Carroll</i>	73
Challenges in Mining Social Media: Sparsity and Quality <i>Thomas Gottron</i>	73
Simulation Data Mining in Artificially Generated Data <i>Steven Burrows</i>	74
Unsupervised entailment detection for IE query expansion <i>Ted Briscoe</i>	74
Piggyback: Using Search Engines for Robust Cross-Domain Named Entity Recognition <i>Hinrich Schütze, Massimiliano Ciaramita</i>	74
Feedback Methods in Large Scale Visual Document Analysis <i>Michael Granitzer</i>	75
The Optimum Clustering Framework <i>Norbert Fuhr</i>	75
TIRA – Research Assistant for Empirical Evaluations <i>Tim Gollub</i>	76
LFRP-Search: Multi-Layer Faceted Search with Ranking and Parallel Coordinates – Interactive Retrieval for Complex Documents and Individual Information Needs <i>Andreas Henrich</i>	76
Beyond Search: Interactive Web Analytics <i>Alexander Loeser</i>	77
The Retrievability of Documents <i>Leif Azzopardi</i>	77
Looking at Document Collections from a Bird’s Eye View: Exploratory Search as Based on the Context Volatility of Terms <i>Gerhard Heyer</i>	78
Facet-Streams and Search-Tokens – Tangible User Interfaces for Information Seeking <i>Harald Reiterer</i>	78
What do you mean? – Determining the Intent of Keyword Queries on Structured Data <i>Wolf Siberski</i>	79

Exploiting Unlabeled Data in Information Retrieval Tasks <i>Benno Stein</i>	79
Cross-lingual adaptation using structural correspondence learning <i>Peter Prettenhofer</i>	80
Search by Strategy <i>Arjen P. de Vries</i>	80
Challenges in Patent Retrieval and Mining <i>Dennis Hoppe</i>	81
Cancelled Talks	81
The Working Groups	
Apparatus, Reproducibility, and Evaluation	82
What can computers learn from reading one million books?	87
Retrieval of Complex Structures	90
Sentiment and Opinion	93
A Theme: Towards More Open Search In Europe	
Discussion	96
Recap of the Proposal	97
Acknowledgements	98
Participants	99

3 Overview of Talks

3.1 Knowledge Discovery in the Web: Potential, Automation and Limits

Stefan M. Rüger (The Open University, GB)


License  Creative Commons BY-NC-ND 3.0 Unported license
© Stefan M. Rüger

What is wrong with a picture of Tony Blair hunting and pecking the keyboard in the midst of school children staring at their respective screens? What works on the web, and what doesn't? Can we use Linked Open Data to persist and share experimental setups? Stefan's talk reflects on the value and potential of interlinked data, semantic web, social networks and data-mining. At the same time he elicits new research directions, which are only enabled by the sheer mass of data, sensors, facts, reports, opinions and inter-linkage of people.

There are a number of information requests that traditional web services can cover well: shortest distance from A to B, opening times of theatre plays, book reviews given a snapshot of a book cover, extensive and competent wikipedia articles on virtually every aspect of our lives. We can hardly imagine a world that does not offer these web services, which have complemented traditional methods of resource discovery, say in libraries. It is quite possible that automated processing of excessive amounts of data paired with new methods of semantic web, information retrieval, human information interaction, social networks, and cloud computing opens up possible new areas of knowledge discovery: Will we have wikiversities and do we want them? Will activities such as TrueKnowledge and Wolfram Alpha be in a position to answer all complex questions that have factual answers? Will the tradition of linear argumentation in printed scientific articles give way to graphical, linked argumentation landscapes? What would be possible or necessary instruments to answer questions such as "What led to the most recent war in Iraq?" Can we create digital research labs such as virtual microscopes and telescopes? What can we learn through computers being able to learn 1 million books? Or watch the news of thousands of television channels in 100 countries?

3.2 Full Lifecycle Information Extraction and Multiparadigm Indexing with GATE

Hamish Cunningham (University of Sheffield, GB)

License  Creative Commons BY-NC-ND 3.0 Unported license
© Hamish Cunningham

Hamish summarised the last few years work with GATE (<http://gate.ac.uk/>), including developments around collaborative manual annotation, process support tools, cloud computing and a mixed-mode index server: "How I was sentenced to 20 years hard labour and the potential implications for over-priced IR systems":

We talk, we write, we listen or read, and we have such a miraculous facility in all these skills that we rarely remember how hard they are. It is natural, therefore, that a large proportion of what we know of the world is externalised exclusively in textual form. That fraction of our science, technology and art that is codified in databases,

taxonomies, ontologies and the like (let's call this *structured data*) is relatively small. Structured data is, of course, machine tractable in ways that text can never be (at least in advance of a true artificial intelligence, something that recedes as fast as ever over the long-term horizon). Unfortunately structure can also be inflexible and expensive to produce in ways that text is not.

Language is the quintessential product of human cooperation, and this cooperation may well have shaped our capabilities and our culture more than any other single factor. Text is a beautiful gift that projects the language of particular moments in time and place into our collective futures. It is also infernally difficult to process by computer (a measure, perhaps, of the extent of our ignorance regarding human intelligence).

When scientific results are delivered exclusively via textual publication, the process of replicating these results is often inefficient as a consequence. Although advances in computational platforms raise exciting possibilities for increased sharing and reuse of experimental setups and research results, still there is little sign that scientific publication will cease its relentless growth in the near future.

This talk summarises a research programme (now 20 years old) that has resulted in GATE, a General Architecture for Text Engineering (<http://tinyurl.com/gatebook>). In recent years GATE has grown from its roots as a specialist development tool for text processing to become a rather comprehensive ecosystem bringing together software developers, language engineers and research staff from diverse fields. GATE now has a strong claim to cover a uniquely wide range of the lifecycle of text-related systems. It forms a focal point for the integration and reuse of advances that have been made by many people (the majority outside of the authors' own group) who work in text processing for biomedicine and other areas. The talk seeks to draw together a number of strands in this work that are normally kept separate, and in so doing demonstrate that text analysis has matured to become a predictable and robust engineering process. The benefits of deriving structured data from textual sources are now much easier to obtain as a result.

In line with the trends towards openness in life sciences R&D and in publishing, GATE is 100% open source. This brings the usual benefits that have been frequently recognised (vendor independence; security; longevity; flexibility; minimisation of costs; etc.). Less often remarked upon but nonetheless significant in many contexts are traceability and transparency. Findings that are explicable and fully open are often worth much more than results that appear magically (but mysteriously) from black boxes.

[Hamish Cunningham, *The Infernal Beauty of Text*, 2011.]

3.3 Unsupervised and Semi-Supervised Approaches to Cross-Domain Sentiment Classification

John A. Carroll (University of Sussex – Brighton, GB)

License © ⓘ ⊖ Creative Commons BY-NC-ND 3.0 Unported license
© John A. Carroll

Joint work of Carroll, John A.; Read, Jonathon; Zagibalov, Taras; Bollegala, Danushka; Weir, David
Main reference Danushka Bollegala, David Weir and John Carroll, “Using multiple sources to construct a sentiment sensitive thesaurus for cross-domain sentiment classification,” Proc. 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, 2011.
URL <http://www.informatics.sussex.ac.uk/research/groups/nlp/carroll/papers/acl11.pdf>

In recent work we have been addressing the problem of developing accurate sentiment classifiers for domains and languages for which there is little or no appropriate training data.

In one strand of this work, we have developed a sentiment classification method that is applicable when we do not have any labeled data for a target domain but have some labeled data for multiple other domains. The approach is based on automatically creating a sentiment sensitive thesaurus in order to find the association between words that express similar sentiments in different domains. Unlike previous cross-domain sentiment classification methods, our method can efficiently learn from multiple source domains.

In a second strand, we have developed a novel unsupervised sentiment classification technique, based on a ‘seed’ vocabulary of positive/negative words, and iterative retraining to bootstrap a substantial high quality training corpus from unlabelled documents. The technique has been applied successfully to text in Chinese, Japanese, Russian and English.

3.4 Challenges in Mining Social Media: Sparsity and Quality

Thomas Gottron (Universität Koblenz-Landau, DE)

License © ⓘ ⊖ Creative Commons BY-NC-ND 3.0 Unported license
© Thomas Gottron

Joint work of Naveed, Nasir; Gottron, Thomas; Kunegis, Jérôme; Che Alhadi, Arifah
Main reference Nasir Naveed, Thomas Gottron, Jérôme Kunegis and Arifah Che Alhadi, “Bad News Travel Fast: A Content-based Analysis of Interestingness on Twitter,” Proc. 3rd ACM International Conference on Web Science (WebSci’11).
URL http://www.websci11.org/fileadmin/websci/Papers/50_paper.pdf


Online communities generate large amounts of text based contents. These contents, however, are very different under several aspects. They cover a wide variety of topics and languages, differ in style and length and are of very different quality. Especially the short texts in microblogs and their sparsity of features pose challenges for many applications in the fields of text retrieval, classification or clustering.

We enrich the representation of microblog entries by annotating them with second level features, such as topics or sentiments. This richer representation can then be used to derive a notion of content quality for social media. Already by itself, this allows for interesting insights into the dynamics of social media.

Preliminary results further suggest that this notion of content quality can be used as a static quality measure to improve retrieval on microblogs.

3.5 Simulation Data Mining in Artificially Generated Data

Steven Burrows (Bauhaus-Universität Weimar, DE)


License  Creative Commons BY-NC-ND 3.0 Unported license
© Steven Burrows

Systems simulation and data mining can complement one another to form simulation data mining and harness intelligence to automatically suggest improvements to simulation models. One challenge in simulation data mining is to develop recommendations for competing variables. In aviation, for example, simulation data mining has been applied to develop cost-effective compromises between flight time and maintenance efforts over aircraft lifetimes.

This talk will introduce a simulation data mining project called Matilda (Mining Artificial Data). In this project, a broader view of document mining is taken to consider artificially generated non-text documents containing data for bridge specifications. It will be shown how simulation data mining can be applied to support the interactive design of bridge structures by removing some of the work required to manually simulate design iterations in turn. Another area of potential is to apply clustering algorithms to automatically identify related models with the development of appropriate distance measures.

3.6 Unsupervised entailment detection for IE query expansion

Ted Briscoe (University of Cambridge, GB)

License  Creative Commons BY-NC-ND 3.0 Unported license
© Ted Briscoe

Main reference BioNLP 2011, A workshop of ACL/HLT 2011, Portland, Oregon, USA
URL <http://compbio.ucdenver.edu/BioNLP2011/program.shtml>

Query expansion for IE systems hasn't been explored much, but would be useful in contexts where, say, the user identified a prototypical predicate denoting a relation of interest. The first step in such an approach would be to detect entailed predicates given the user-defined predicate.

Entailment detection systems are generally designed to work either on single words, relations or full sentences. We develop a new approach – detecting entailment between dependency graph fragments of any type – which relaxes these restrictions and leads to much wider entailment discovery. An unsupervised framework is described that uses intrinsic similarity, multi-level extrinsic similarity and the detection of negation and hedged language to assign a confidence score to entailment relations between two fragments.

3.7 Piggyback: Using Search Engines for Robust Cross-Domain Named Entity Recognition

Hinrich Schütze (Universität Stuttgart, DE), Massimiliano Ciaramita (Google Zurich)

License  Creative Commons BY-NC-ND 3.0 Unported license
© Hinrich Schütze, Massimiliano Ciaramita


Joint work of Schütze, Hinrich; Ciaramita, Massimiliano
URL <http://ifnlp.org/schuetze/piggyback11/piggyback11.pdf>

We use search engine results to address a particularly difficult cross-domain language processing task, the adaptation of named entity recognition (NER) from news text to web

queries. The key novelty of the method is that we submit a token with context to a search engine and use similar contexts in the search results as additional information for correctly classifying the token. We achieve strong gains in NER performance on news, in-domain and out-of-domain, and on web queries.

3.8 Feedback Methods in Large Scale Visual Document Analysis





Michael Granitzer (Know-Center Graz, AT)

License     Creative Commons BY-NC-ND 3.0 Unported license
© Michael Granitzer

Michael discusses feedback methods utilizing visual representations of large document sets in order to adapt algorithmic parameters and metrics; in this regard, he presents a system for visualizing and analyzing topical, temporal and metadata-based correlations in large document sets. The methods combine well-known visualization techniques for large document sets (i.e. multi-dimensional scaling, self-organized maps) with recent techniques to learn high-dimensional metrics. His talk focuses on two methods: (1) a user can directly manipulate the parameters of the algorithm generating the visualization, and (2) a user can interactive label visualized elements. For getting an impression of the content of a document, Michael shows that the effectiveness of key phrases are as good as text summaries, but much more efficient.

3.9 The Optimum Clustering Framework

Norbert Fuhr (Universität Duisberg-Essen, DE)

License     Creative Commons BY-NC-ND 3.0 Unported license
© Norbert Fuhr


Starting point are the following questions: Is there a principle similar to the Probability Ranking Principle for document clustering? Can we define a cluster metric in which quality is relative to shared relevance to a set of queries?

To answer these (and related questions) Norbert introduces a theoretic foundation for optimum document clustering. Key idea is to base cluster analysis and evaluation on a set of queries, by defining documents as being similar if they are relevant to the same queries. Three components are essential within this optimum clustering framework (OCF): (1) a set of queries, (2) a probabilistic retrieval method, and (3) a document similarity metric.

Based on these components appropriate validity measure can be introduced, and optimum clustering can be defined with respect to the estimates of the relevance probability for the query-document pairs under consideration. Norbert shows that well-known clustering methods are implicitly based on the three components, but that they use heuristic design decisions for some of them. Altogether, the OCF can help to make targeted research for developing better document clustering methods possible.

3.10 TIRA – Research Assistant for Empirical Evaluations

Tim Gollub (Bauhaus-Universität Weimar, DE)

License  Creative Commons BY-NC-ND 3.0 Unported license
© Tim Gollub

Joint work of Gollub, Tim; Stein, Benno


We present and demonstrate TIRA, software that supports the data mining community in conducting replicable and comparable experimental analysis.

Although desired, the comparison of empirical evaluations in scientific publications is often impossible due to differing datasets, baselines, or parameter settings. Furthermore, details concerning the implementation of algorithms are not given in many cases. With TIRA, we want to contribute to a more efficient and cooperative research community that benefits from sharing data, algorithms, experiments, and CPUs.

TIRA provides an open framework for designing, running, and publishing data mining experiments. We try to acquire a comprehensive list of datasets, algorithms, and evaluation metrics. TIRA manages parametrization and distribution of experiments, as well as storing results for further performance studies and visualizations.

3.11 LFRP-Search: Multi-Layer Faceted Search with Ranking and Parallel Coordinates – Interactive Retrieval for Complex Documents and Individual Information Needs

Andreas Henrich (Universität Bamberg, DE)


License  Creative Commons BY-NC-ND 3.0 Unported license
© Andreas Henrich

In enterprise search scenarios information needs and documents are quite diverse. In addition, information needs often are vague and unclear, which means that users cannot explicitly define the search criteria that specify their search request. In these cases, exploratory search approaches are necessary to support users in interactively refining their search queries.

To address such settings, Andreas presents a retrieval system for complex search situations that is based on four constituent parts. The approach deals with the heterogeneity of potential target objects when performing a search considering multiple artefact layers (e.g., projects, products, persons, and documents). The overall system is designed as a kind of data warehouse which supports relations between artifact types and considers them in ranked retrieval. The biggest effort for building such a system is the specification of the ETL (extract / transform / load) processes for the different sources. To cope with result sets of different granularity, ranking facilities based on facet values as well as Query-by-Example functionalities are included. Parallel coordinates are used to visualize the characteristics and dependencies of (intermediate) results in order to provide users with a deeper understanding of the data under investigation. In his talk, Andreas discussed the conflicting priorities of ease of use and expressive power.

3.12 Beyond Search: Interactive Web Analytics

Alexander Loeser (TU Berlin, DE)


License  Creative Commons BY-NC-ND 3.0 Unported license
© Alexander Loeser

Today, the Web is one of the world's largest databases. However, due to its textual nature, aggregating and analyzing textual data from the Web analogue to a data warehouse is a difficult problem. For instance, users may start from huge amounts of textual data and drill down into tiny sets of specific factual data, may manipulate or share atomic facts, and may repeat this process in an iterative fashion.

Andreas presents the GooLap System (<http://www.goolap.info/>) for interactive fact retrieval from the Web supporting several dozens of named entity types and the corresponding relationships. The keyword-based query interface focuses on both, simple query intentions, such as, “display everything about Airbus” or even complex aggregation intentions, such as “List and compare mergers, acquisitions, competitors and products of airplane technology vendors.” For retrieval, he describes a new query language which also allows for joins on the extracted fact relations. Andreas discusses the fundamental problems in the iterative fact retrieval process: What are common analysis operations of “business users” on natural language Web text? What is the typical iterative process for generating, verifying and sharing factual information from plain Web text? Can we integrate both, the “cloud”, a cluster of massively parallel working machines, and the “crowd”, such as users of GoOLAP.info, for training 10.000s of fact extractors, for verifying billions of atomic facts or for even generating analytical reports from the Web?

3.13 The Retrievability of Documents

Leif Azzopardi (University of Glasgow, GB)


License  Creative Commons BY-NC-ND 3.0 Unported license
© Leif Azzopardi

How do individuals interact with information? Can we exploit models from transportation and planning to derive new suite of measures for information retrieval?

Leif introduces the concept of accessibility from the field of transportation planning and explains how it can be adopted within the context of information retrieval. By drawing analogy between the two fields we are able to develop a new suite of measures for information retrieval that considers how easily a document can be retrieved using a particular retrieval system. This intuitive measure provides the basis for examining different aspects of the influence of a retrieval system has upon the documents in the document collection. He shows that distribution of retrievability values over a document collection depends on the employed retrieval model and discusses possible ways in which retrievability measures might be used for document mining tasks.

3.14 Looking at Document Collections from a Bird’s Eye View: Exploratory Search as Based on the Context Volatility of Terms

Gerhard Heyer (Universität Leipzig, DE)


License  Creative Commons BY-NC-ND 3.0 Unported license
© Gerhard Heyer

Was is exploratory search? In many text retrieval applications the user is not primarily interested in facts, but more so in the way facts are reported on, and how the conceptualizations and judgments expressed in the reporting texts are changing over time. Investigations in journalism, technology mining, or eHumanities are examples that can be considered as instances of exploratory search. The classical paradigm of query-and-retrieve is not suitable here because a user who wants to retrieve those documents that best satisfy his information needs first needs support (a) to make himself familiar with a search domain, (b) to identify terms that are of potential interest to the topic he is researching, and (c) to follow variant paths to explore the domain of interest.

Gerhard presents the notion of context volatility of terms as a measure for the interest-iness of terms. Basically, this measure computes the variance of a term’s co-occurrences during some period of time. Given a time-stamped collection of documents, terms with a high degree of context volatility for some period of time usually represent “hot topics” as they have been controversially discussed during that period. Using this measure of context volatility, he presents a system for exploratory search that for a time-stamped collection of documents, such as the New York Times corpus, (1) generates a list of “hot topics”, and (2) supports the user to identify those periods of time where the discussion of these “hot topics” erupts. The whole search process is highly interactive, and an instructive instance of visual analytics.

3.15 Facet-Streams and Search-Tokens – Tangible User Interfaces for Information Seeking

Harald Reiterer (Universität Konstanz, DE)

License  Creative Commons BY-NC-ND 3.0 Unported license
© Harald Reiterer

Social activities such as collaborative work and group negotiation can be an essential part of information seeking processes. However, they are not sufficiently supported by today’s information systems as they focus on individual users working with PCs. Reality-based UIs with their increased emphasis on social, tangible, and surface computing have the potential to tackle this problem. By blending characteristics of real-world interaction and social qualities with the advantages of virtual computer systems, they inherently change the possibilities for collaboration, but until now this phenomenon has not been explored sufficiently. Harald presents two examples of Tangible User Interfaces (TUIs) and analyzes their power and expressiveness for information seeking activities.


The first example is “Facet-Streams”, a hybrid interactive surface for co-located collaborative product search on a tabletop. Facet-Streams combine techniques of information visualization with tangible and multi-touch interaction to materialize collaborative search on a tabletop. It harnesses the expressive power of facets and Boolean logic without exposing users to complex formal notations. The second example is “Search-Tokens”, also a hybrid

interactive surface for co-located collaborative information exploration on a tabletop. The physical appearance of the Search-Tokens provides a higher visual and tangible affordance than a GUI that is solely based on digital sliders, text fields, buttons, etc. By placing a Search-Token on the tabletop, it is augmented by visualizations, when a search criterion is entered, rotating the Search-Token allows users to define the criterion's weight.

But, how useful are these new user interaction paradigms? Can search be made more like interacting with the non-digital world? Harald reports on user studies that reveal how visual and tangible expressivity are unified with simplicity in interaction, or how different strategies and collaboration styles are supported.

3.16 What do you mean? – Determining the Intent of Keyword Queries on Structured Data

Wolf Siberski (Universität Hannover, DE)

License  Creative Commons BY-NC-ND 3.0 Unported license
© Wolf Siberski


More and more factual information is mined from the Web and provided as structured data, for example in the Linked Data initiative. While this opens up the potential for fulfilling information needs much better than just Web page content, it also requires new approaches to retrieval, because relevance functions for text don't work well on structured data, and complex query languages are the wrong tool for most users.

As one such novel approach, we have developed QUICK, which combines the convenience of keyword search with the expressiveness of structured queries.

Users start with a keyword query and then are guided through a process of incremental refinement steps to specify the intention of their query. We show how QUICK identifies possible intentions and computes a construction wizard for query intent specification with a few mouse clicks.

3.17 Exploiting Unlabeled Data in Information Retrieval Tasks

Benno Stein (Bauhaus-Universität Weimar, DE)

License  Creative Commons BY-NC-ND 3.0 Unported license
© Benno Stein


Joint work of Stein, Benno; Nedim, Lipka; Meyer zu Eissen, Sven; Prettenhofer, Peter

Data that is labeled with class information is a valuable and expensive resource, and there are different efforts to combine labeled data with unlabeled data to improve classifier effectiveness during machine learning. We discuss three approaches, namely constrained clustering, co-training, and structural correspondence learning, since they pursue quite different data exploitation paradigms. All approaches can be called semi-supervised, whereas constrained clustering exploits additional labeled data in order to inform an - usually - unsupervised analysis. By contrast, co-training as well as structural correspondence learning exploit additional unlabeled data. With respect to the former, we explain why it is difficult to scale-up the co-training idea, i.e., the extension of the training set. With respect to the latter we introduce a cross-language sentiment classification approach where the use of unlabeled

data scales extremely well, leading to a significant improvement over state-of-the-art machine translation baselines.

3.18 Cross-lingual adaptation using structural correspondence learning

Peter Prettenhofer (TU Graz, AT)

License  Creative Commons BY-NC-ND 3.0 Unported license
© Peter Prettenhofer

Joint work of Prettenhofer, Peter; Stein, Benno

Main reference Peter Prettenhofer and Benno Stein, “Cross-lingual adaptation using structural correspondence learning,” *ACM Transactions on Intelligent Systems and Technology* (to appear), ACM, 2011

Cross-lingual adaptation is a special case of domain adaptation and refers to the transfer of classification knowledge between two languages. We describe an extension of Structural Correspondence Learning (SCL), a recently proposed algorithm for domain adaptation, for cross-lingual adaptation in the context of text classification. The proposed method uses unlabeled documents from both languages, along with a word translation oracle, to induce a cross-lingual representation that enables the transfer of classification knowledge from the source to the target language. The main advantages of this method over existing methods are resource efficiency and task specificity.


We report on experiments in the area of cross-language topic and sentiment classification involving English as source language and German, French, and Japanese as target languages. The results show a significant improvement of the proposed method over a machine translation baseline, reducing the relative error due to cross-lingual adaptation by an average of 30% (topic classification) and 59% (sentiment classification). We further report on empirical analyses that reveal insights into the use of unlabeled data, the sensitivity with respect to important hyperparameters, and the nature of the induced cross-lingual word correspondences.

References

- 1 Peter Prettenhofer and Benno Stein. *Cross-lingual adaptation using structural correspondence learning*. *ACM Transactions on Intelligent Systems and Technology* (to appear), 2011
- 2 Peter Prettenhofer and Benno Stein. *Cross-language text classification using structural correspondence learning*. In *Proceedings of the 48th Annual Meeting of the Association of Computational Linguistics*, Uppsala, Sweden, 2010

3.19 Search by Strategy

Arjen P. de Vries (CWI Amsterdam, NL)

License  Creative Commons BY-NC-ND 3.0 Unported license
© Arjen P. de Vries


Today Dagstuhl, tomorrow the world. Making IR and DB engineers redundant with a dynamically reconfigurable DB-based IR system (<http://devel.spinque.com/BilthovenDemo/>). For interactive information access Arjen and his group has been developing a system that allows for visually constructing a search strategy by connecting building blocks. The backend is a combined IR-DB system based on probabilistic relational algebra.

The topic of his talk is focused on the process after text mining has taken place – assuming that the mining process would lead to semantic annotations of the original document space.

The key idea is to let users decide for themselves how to navigate this “semantically” enriched document space, but then do support them in exploration of this annotated data. Hereto, a new interaction paradigm is followed, called “search by strategy”: an iterative two-stage search process that separates search strategy definition (the “how”) from the actual searching and browsing of the collection (the “what”). A visual query environment (the “search strategy builder”) allows professional searchers to visually express their search strategy for a particular need, and then execute this strategy on a probabilistic relational database system. The idea is that by bringing together the visual search strategy definition and faceted browsing of result sets will allow the user to discover during query formulation which semantic annotations are useful for their information need, and exploit their value for finding better search results in less time.

3.20 Challenges in Patent Retrieval and Mining

Dennis Hoppe (Bauhaus-Universität Weimar, DE)

License  Creative Commons BY-NC-ND 3.0 Unported license
© Dennis Hoppe

Over 50 million multi-language patents no longer permit that we “search as we did 28 years ago,” answered Henk Tomas in 2007 at the IRF Symposium as he was asked about the greatest problem of patent retrieval. Patent retrieval concerns not only companies and inventors, but also occasional users and researchers.

Current challenges in patent retrieval include the need for multi-language support, robustness against errors in patent text caused by error-prone OCR recognition, handling inconsistencies in metadata, and yielding a high recall while retrieving patents. However, various issues make the retrievability of patents difficult: (1) domain-specific vocabulary, (2) complex syntactic structure, (3) vague translations of companies and persons from far-east, and (4) companies change their names over time, merge or demerge.

The emphasis of this talk is on methods to improve patent retrievability such as automated patent classification, and image-based patent retrieval and classification.

3.21 Cancelled Talks

A few participants were unable to present their talks due to unforeseen circumstances:

- C. J. Keith van Rijsbergen (University of Cambridge) Title: Document Clustering Revisited
- Ingo Frommholz (University of Bedfordshire) Title: A Quantum-inspired Polyrepresentation Framework
- Oliver Niggemann (Hochschule Ostwestfalen-Lippe) Title: A Probabilistic MajorClust Variant
- Marc Lechtenfeld (Universität Duisburg-Essen) Title: Determining the Polarity of Postings for Discussion Search
- Oren Etzioni (University of Washington)

4 The Working Groups

The following working groups were proposed by the participants:

- Out-takes: negative results in information retrieval.
- Conference reviewing mechanisms.
- Apparatus, setups, infrastructure for reproducibility.
- Human-in-the-loop scenarios, evaluation.
- Automated readers: knowledge extraction from on-line million-book libraries.
- Retrieval of complex structures.
- Sentiment, opinion, and social media.
- Optimum clustering.

From these suggestions we selected the following:

1. Apparatus, setups, infrastructure for reproducibility, plus Human-in-the-loop scenarios, evaluation.
2. Reading 10 million books. Automated readers: knowledge extraction from on-line million-book libraries.
3. Retrieval of complex structures. Cross-over of data-oriented methods and retrieval methods: models of structured data retrieval (e.g. entities, entity graphs).
4. Sentiment, opinion, and social media.

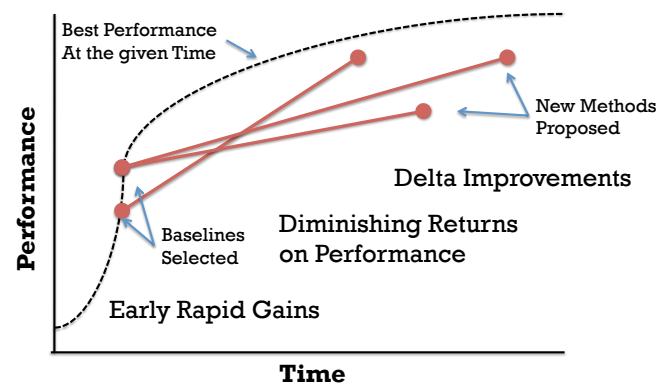
4.1 Apparatus, Reproducibility, and Evaluation

Rapporteur: Leif Azzopardi

4.1.1 Building Research Infrastructure

This working group considered the design and development of common frameworks, applications and resources for experimentation in Information Retrieval (IR) and Document Mining (DM). The group considered a number of aspects on this topic: (1) the current problems with evaluations and comparisons of methods, models and systems, (2) the components and building blocks of experimental research, the research process and the infrastructure, and (3) the ideal research infrastructure, its benefits, drawbacks, barriers and challenges.

Initial discussions acknowledged that many of the experiments and empirical work that is currently undertaken is performed independently and often under very different conditions. Each research group and potentially each researcher designs and develops experimental scripts and processes to conduct experimental research to determine the effectiveness and efficiency of methods, models and systems. All the possible variables and factors in the experimental process (such as the algorithm implementations, the choice of the algorithms/methods used, the toolkits used, the data processing, the parameter tuning, etc) vary across the body of research reported in the field. In conjunction with the pressure to publish results that show "significant" differences, this can lead to overly optimistic results being reported. Subsequently, published work often contains comparisons where poor baselines have been used, newly proposed methods are over trained, or some other manipulation to help bolster the evidence for the method (to try and maximise the chance of publication). The consequence of this is that many of the experiments and the results reports and methods described in the literature are difficult to replicate and validate. Also, it is difficult to contextualize any performance improvements made by different retrieval and mining methods across the research area.

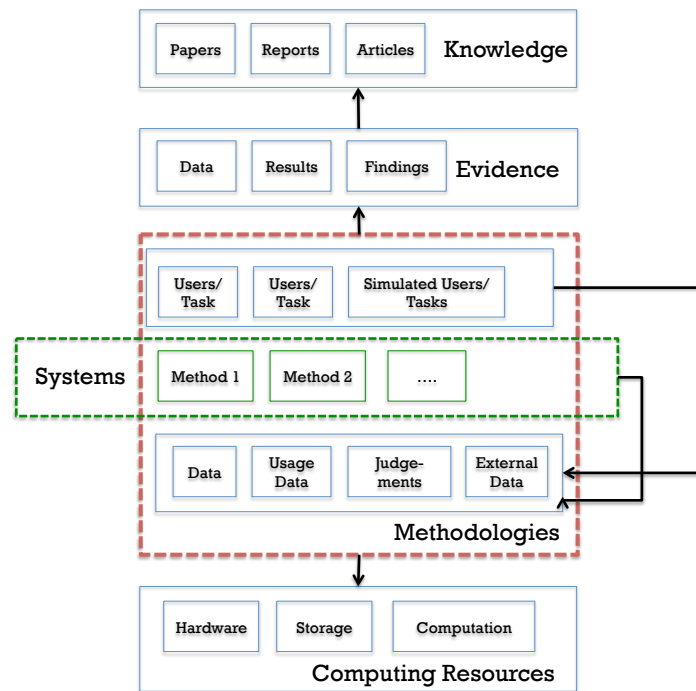


■ **Figure 1** The Armstrong Effect: The current best performance is rarely used when proposing a new method. Instead, an older baseline is preferred and used. The dotted line represented a smoothed version of the overall progress in the field in terms of performance for a given task.

For example, in Amstrong et al [?], they showed that evaluations conducted within IR rarely used the current state of the art to perform comparisons. Instead, older and weaker baselines were used in order to compare the proposed methods. As a result, newly proposed models would significantly outperform the weak baselines, yet it would rarely outperform the current state of the art. In Figure 1, we have illustrated this effect, where the dotted black line denotes the best performance at the time (smoothed across innovations), and the lines joined by balls denote a baseline vs. a proposed method. The shape of the dotted black line shows that initially there is usually early rapid gains in performance when a task is introduced. Then diminishing returns in performance begins to kick in, before the work becomes delta (i.e. only marginal improvements are being achieved by the best newly proposed methods. If the actual performance across the field was known then the progress being made on a particular tasks could then be readily contextualized and compared accurately.

Otherwise, without a centralized repository of results and access to standardized configurations it is difficult to determine whether significant progress is actually being made in an area or not. i.e. have we reached the top of an innovation curve/cycle with the current methods and technology? Also, currently, reviewers must invest a lot of trust in the work presented to them, and they have to assume that the reported results hold (and that the experiments and findings are repeatable and reproducible by others. For those replicating existing methods, they face a gambit in that, if they discover that they can not reproduce findings similar to past work, or worse they obtain contradictory finding, then they run the risk of trying to publish negative findings, which was perceived to be almost impossible. This means many researchers waste a lot of time discovering certain methods do not perform as well as purported, leaving them in a difficult position. Essentially, there was a consensus among participants in the working group that the current way we perform research leaves a lot to be desired. And a potential solution would be to provide some sort of common evaluation infrastructure, which extends and enhances current evaluation forums like TREC, CLEF, etc. The infrastructure would provide a standardized environment where:

1. experiments are reproducible and repeatable,
2. verification of results is performed externally and independently,
3. the cost of experimentation is significantly reduced, and
4. data from previous experiments and methods would be available for comparison purposes.



■ **Figure 2** An abstraction of the research process and supporting research infrastructure.

It was hypothesized that if a common framework for experimentation existed which would enable highly controlled evaluation, then the above problems may be mitigated. While, it would provide a number of other benefits, it was also acknowledged that there is likely to be a number of potential problems in designing and developing such infrastructure.

To discuss this hypothesized framework, we took a very abstracted view of what such infrastructure might be, and assumed task independence. i.e. we imagined that the framework could be applied to various retrieval and mining tasks, like ad-hoc retrieval, classification, clustering, etc. The purpose of such a framework would be to:

- enable the evaluation of systems, methods and models,
- ensure replication and repeatability of experiments,
- facilitate the dissemination and reporting of research results,
- improve experimenter efficiency by lowering the start up costs of running and replicating experiments,
- focus research on a particular well defined task, and
- provide standardization, automation and controlled experimentation.

4.1.2 Towards an Ideal Research Infrastructure

Figure 2 provides a high-level schematic of our perspective on the research process and how the research infrastructure differs from existing evaluation forums and tools. The end goal of the research process is to generate new knowledge, to question and verify existing knowledge and to revise knowledge. These conclusions and implications are supported by the experimental data collected, the results obtained and the findings made. To create this experimental data a scientific methodology is employed, and in IR and DM, usually consists of a several key components:

- Different sets of users (even simulated users) and corresponding tasks, potentially along with their interactions.
- A system (or set of sets), such as Lemur or Terrier, to perform the retrieval or mining with, and a series of different methods/models that are to be tested.
- Data to form the test collections for evaluation, the judgments, external data and the measurements that are to be used.

Put together a set of these components creates a methodology. Evaluation forums like TREC and CLEF often define the process and the components, and let the systems and methods be varied. The research infrastructure would provide a system or an interface for systems to enable the testing of methods (or combinations of methods), house a repository of data and results. The ideal infrastructure would be able to cater for various users, including researchers, academics, businesses, living labs and organizations that facilitate research. It was acknowledged that different users will have their own different needs, concerns, constraints and problems. But they should also be able to benefit from the infrastructure as well.

Collaboration, Cooperation and Competition in Research: Having an abstraction of the ideal infrastructure, then promoted discussion in the working group on how collaboration and cooperation between researchers and research groups might improve. It was mooted that, the research infrastructure would be able to promote sharing and trust between researchers. And this would create a cooperative environment where science of IR is driven forward in a transparent manner. The research infrastructure would also help to keep researchers honest, because all the results and findings are independently verified and recorded by the infrastructure. Though the collaboration and cooperative it would be possible to perform reproducible and repeatable experiments. So while, *Competition drives Innovation* it can potentially obfuscate and diffuse the progress of the research within the field. Through cooperation the field may be able to identify very quickly areas of research where increases in performance are marginal and becoming delta.

However, the decision by researchers to cooperate or to compete presents the *The Researcher's Dilemma*. If researchers choose not to cooperate, then we continue with the current state of evaluation within IR and DM, i.e. piecemeal and ad-hoc testing and reporting of results. As mentioned above, the downside is that progress is obfuscated and misleading. However, if researchers cooperate by building an open research infrastructure, then there is the potential to advance progress in the area rapidly, and to identify when progress in a particular topic is converging (leading to only marginal or delta improvements). Of course, there may be researchers that cooperate, and others that stay in competition. Here, the cooperative researchers, run the risk of losing out to the competitors. The cooperative researchers are likely to forgo obtaining short term gains through publications, because they are investing in the development of the infrastructure. If the infrastructure fails to be adopted or does not obtain critical mass then the cooperative researchers will incur even greater losses. The competing researchers would be able to gain the upper hand, by continuing to publish without having to conform to any standard. For a competing researcher to switch and become a cooperative researcher is also a difficult choice, as they may already have custom built infrastructure which is giving them a strategic advantage over other researchers. Without some external regulation (if reviewers demanding independent verifications, for instance, or funding agencies requiring publicly accessible results), then the cost of switching is likely to be too high. Given the publish or perish environment, this will potentially provide the competing researchers to stay in the game, while unsuccessful cooperative researchers will need to find work elsewhere (in particular for early researchers). An alternative is if the

cooperative forms a closed piece of research infrastructure, where only those participating have access. This may provide enough protection for cooperative researchers in the short term to mitigate the risks. Nonetheless, researchers face a decision in how to act and work. At times, we compete and other times we cooperate. In terms of building research infrastructure, a clear case for the different positions and what can be gained through cooperation needs to be put together.

4.1.3 Challenges and Barriers

Barriers to Adoption: While the idea of having research infrastructure in place was generally viewed as a positive direction for the field, we also recognized that there are a number of barriers to its adoption. These included:

- the framework/infrastructure may impose too many constraints
- if the infrastructure is difficult to use or requires a steep learning curve
- the cost of setting up experiments in the new infrastructure might too high, and not worth the switch, and
- researchers may be unwilling to give up any strategic advantage that they currently have with their own custom-built infrastructure.
- researchers may not want to participate because their methods may not achieve the same performances that they have either reported and/or produced under their own conditions.

Success Factors: For the success of a common research infrastructure, the group identified a number of key factors which would help its adoption, and provide benefit to its users. These were:

- Getting a critical mass of participants, in particular the major research groups in the field
- Providing a flexible and general interface that can support the various kinds of research methods to be tested
- Providing a reasonable good level of support, documentation and ensuring that the infrastructure is easy to use
- Ensuring engagement and usage through forums and workshops (such as TREC, CLEF, etc)
- Providing transparency in the research process
- Lowering the cost of research, and lowering the entry/start up costs to research
- Valuing of the results and findings from the research infrastructure by reviewers over those from out-with such infrastructure, since the research infrastructure will be providing independent verification of the results.

Further Challenges: The group mainly discussed the issues associated with the design and development of research infrastructure such that it would support experimental research in IR and Data Mining. A major challenge is to operationalize this vision and to produce a concrete design of the infrastructure, and then develop it. Then, further challenges will emerge in terms of sustainability and maintainability. To address these resources will be needed to create and then sustain the initiative. However, with the development of such infrastructure likely to be quite costly, requiring time, money and manpower, it is likely that such an initiative would require a long-term project to be funded.

4.2 What can computers learn from reading one million books?

This working group took up Greg Crane’s question “What Do You Do with a Million Books?” (D-Lib Magazine, 12(3), March 2006) and contextualised it to the Dagstuhl workshop theme of document mining.

An active reading life of 70 years enables us to read around 25,000 books in a lifetime assuming one can finish one book per day on average. Ongoing digitisation efforts have, of course, created many more books in computer-readable form. So, what could we mere humans learn from computers that read one million books for them? And what does it mean that a computer reads a book, for that matter?

To give a simple example, one hope is that algorithms are able to assign one or more topics to passages of text. These topics are embedded in the context of other topics and sentiment (see also the workgroup on sentiment and opinion in Section 4.4) as well as being spread over place and time. This alone should provide for interesting automated analysis how the topic of, for example, slavery, has shifted over time, with different speed in the UK and the US, where slavery was abolished much later.

Our working group revolved around developing a set of challenges and then exploring possible methods and approaches to tackle them.

4.2.1 Challenges v1.0

We discussed a range of challenges and activities for computer algorithms; here is our summary of key challenges roughly in ascending order of difficulty:

- Identify key phrases — already happening in any self-respecting natural language processing (NLP) package
- Detect topics — a more difficult task, as it involves document understanding, but already currently tackled more or less successfully by the NLP community
- Discover memes or ideas — involves yet another level of abstraction
- Visualise peculiarities and outliers given a topic, for example, by creating a bird’s eye view of a topic
- Establish relations, trends and geographic patterns of key phrases, topics, and ideas
- Follow the provenance, propagation and dispersion of ideas
- Compute sentiment associations with topics, ideas and text passages
- Classify documents as novels, scientific papers, recipes, advertisement, travel reports, poems, newspaper articles, school books, essays, manuals, magazine articles, or similarly for images, classify these as landscape scenes, portraits, chemical structures, electrical circuits, mathematical formulae, architectural drawings etc
- Extract summaries of individual documents by assigning importance to sentences and keeping the highest rated ones
- Bias summaries on queries reflecting the information need of the user
- Elicit change of meaning of words, topics or ideas over time (historic semantics), and as corollary, answer the question which words and topics are invariant
- Recommend related material
- Spot prototypical ideas and creativity
- Recognise important documents (original or influential)
- Create a database of facts to be used for question answering tasks
- Enable focussed retrieval for facets and references — with particular focus on the long tail
- Delineate variable quality of writing and writing style

- Generate multi-document summaries
- Compare topics across different languages
- Assign authorship
- Uncover contradictions, lies or polemic
- Describe the knowledge and world view of a period — lending support to the discipline of history of science
- Explore limitations of the world view of a time (including those of the currently held world view!)

These challenges are all connected to the overriding aim to understand documents through machine processing, and some of these assume that there are sufficiently many documents such that a meaningful analysis can be carried out in time windows and that the results inform about (continuous) changes in time.

4.2.2 Challenges v2.0

The next level of challenges would be to attempt an automated *understanding of language* rather than a particular document, broadly assuming that the first level of challenges will have been mastered in the foreseeable future to sufficient quality. Automated language understanding will have been achieved if, for example, computers can

- Generate novels or poetry automatically
- Establish and deploy models of creativity
- Tell apart irony from parody and facts from fiction
- Read aloud and correctly intonate books
- Interpret graphs, figures and diagrams — for example electrical-engineering circuits or chemical structures
- Predict differences in cultures
- Analyse patents as to prior art

4.2.3 Methods

The discussions revolved around the current best practice for attempting some of the challenges that we identified, with a less expressed speculation on what could be used to tackle the more difficult or frankly at this point utopian challenges.

Statistics. A simple counting exercise can yield amazing insights, particularly in very large datasets. For example, a count of the references to locations on the globe in the English-language wikipedia can reveal the areas of the world that are of particular interest to the English-speaking population, ie, yield their “view of the world geography”.

Co-occurrence analysis of words can give key insights, too. For example, terms that occur together and have a similar profile of frequency and volatility may be good candidates for synonyms.

Statistical methods are not confined to counting. They subsume probabilistic language models, the use of ontologies, factor analysis and latent semantic indexing, to name but a few. The latter have shown to be able to crystallise memes, ideas and topics in documents, which in turn are the basis of semantic profiling and many a similarity function between document parts.

Citation analysis and bibliometrics. The overwhelming majority of books and text documents do not feature the explicit references that are common in scientific papers. However, the hypothesis in our discussion was that there is a weak variant of citations by

creating links between books and, at higher granularity, between chapters and passages in the collection through common references to events, situations, ideas or through explicitly copied or cited passages. This gives rise to a mathematical graph from these links that is akin to a citation network. The analysis of such a graph might provide insights into which texts have been original or influential or both (tackling questions such as “is Magna Carta really the cornerstone of the idea of democracy?”)

Methods from Natural Language Processing and Linguistics. NLP and linguistics are the obvious areas from which methods can be derived that might help addressing above challenges. In particular, sentiment analysis is a branch that holds the promise of being able to classify documents as to their perception by readers, while, for example, the branch of forensic linguistics contains some of the ingredients for assigning authorship of documents through internal features or for extracting “fingerprints” of documents — it may even turn out that the analysis of stop word usage alone is a significant contributor.

Visualisation. As with all data mining problems and exploratory statistics, bespoke visualisation delivers important tools for understanding and detecting structures. It was recognised that time plays a vital role in many of the research challenges. We discussed several paradigms, such as a geological metaphor making time the depth of sediments in the earth crust, bespoke animations, a series of snapshots and timeline visualisations. Other ideas focussed on the notion of relevance and information content of document parts that led to the paradigm of islands, ie, relevant text components, emerging from a sea of irrelevance in reaction to a relevance slider that sets a desired threshold. Theme rivers were yet another paradigm touched upon.

Methods from Machine Learning and Data Mining. As with NLP and linguistics, machine learning and data mining are obvious areas that harbour methods to derive aspects of automated language and document understanding:

Any supervised methods such as classification of text passages need ground truth, which may be difficult to obtain and — once obtained — still poses barriers such as poor assessor agreements caused by genuine disparate opinions of the people involved. Crowd sourcing via “games for a purpose” may be a useful tool, in particular with document collections that are of general interest and are likely attract many users. Clustering as an unsupervised method lends itself to visualisation, for example, through Kohonen’s self-organising maps that aim to keep topological neighbourhood relations. Shopping basket analysis and association rules might provide some insight into relations of ideas and topics together with powerful visualisation methods that were developed for this.

4.2.4 Summary

Concluding our discussion, after having reflected on it, there was a real sense that there is a great deal of knowledge that can be discovered from automated document processing. The extent and scale of many and large repositories is likely to support knowledge discovery through redundancy and many examples rather than posing an unwelcome challenge. Knowledge discovery in large text repositories (or on the web) is likely to become a thriving and intellectually challenging research direction.

And what to do when all computational methods have been exhausted for a particular task? As one of the participants remarked: “Well, there is always the option to read for oneself!”

4.3 Retrieval of Complex Structures

Rapporteur: Wolf Siberski

Text document retrieval has reached a high quality standard, and it has become part in many areas of our everyday life (e.g. Web search, library catalogues). In comparison, retrieval of complex structures (RCS) is still in its infancy. The topic has gained interest for several reasons:

- The Semantic Web or Web of Data gets momentum, with hundreds of structured data sources available already as part of the LinkedOpenData initiative
- With advances of entity recognition, entities and their relation can be extracted from documents and provided as graph structure
- A huge amount of structured data is created automatically (e.g. by sensor networks), especially as result of scientific experiments. For all these cases, facilities to search for specific information and/or to explore the complex information spaces are needed.

The goal of the working group was to discuss the nature of retrieval of complex structures (in particular its differences to traditional document retrieval), to discuss the state of research in the field, and to identify promising research directions for this challenge.

4.3.1 Towards a Classification of Complex Structures in IR

The complexity of an information space can have different sources:

- Inherent complexity of the underlying structure (e.g., the database schema).
- Complexity arising from the content type (e.g., CAD models, movies).
- Complexity due to contextual dimensions which are part of the information (e.g., temporal or spatial context).
- Complexity due to the use of different language levels (object language, meta language) or due to the relation between information on different abstraction levels (e.g., the relation between extracted facts and to the documents on which these facts are based).
- Complexity arising from matching different levels of semantic to each other (e.g., IR has to match ambiguous user queries to a structured information space like databases or graphs).
- Complexity arising from integration of different information sources.

The working group identified different “complexity classes” for these complex information spaces, inspired by Chomsky’s formal language hierarchy. In the following, these classes are characterized by the expected or standard means that is necessary to satisfy an information need. The classes are organized from

1. The information need can be satisfied by a bag of words model or one of its variants. One can imagine the information space defined by a set of “flat” documents.
2. The information need can be satisfied by a graph-based data model. In particular, graph structures may be used to specify structured textual entities and their relations (usually represented in a relational schema or an ontology).
3. The information need deals with multimedia aspects of arbitrary kind, which includes also technical drawings.
4. The information need is not restricted, which means this class pertains to information spaces of unlimited structural complexity.

Notice that, in contrast to Chomsky’s model, these classes do not form a hierarchy. Among others, this is rooted in the fact that even highly complex information needs can be formulated as natural language documents, which in turn can be represented as bag of

words models; hence, encodings or transformations are conceivable that losslessly transform information needs across classes. Instead, the complexity classes here are oriented at data models, intended to measure the increasing modeling effort. The proposed class boundaries reflect “natural” modeling leaps as they were experienced by the group members.

4.3.2 Towards a Quantification of Complex Structures in IR

While the introduced “complexity classes” form a qualitative estimate for complex structures in IR, a quantitative estimate is yet missing. As shown by related fields like machine learning, quantifying complexity can yield new theoretical insights on algorithmic properties and achievable accuracy bounds. For example, the Vapnik-Chervonenkis dimension VC [3]—a complexity measure for the richness and expressiveness of hypothesis classes in Machine Learning—induced the development of generalization bounds for different types of classifiers. Similarly, topologies of metric spaces can be measured using different estimates, like for example by the Hausdorff-Besicovitch dimension.

According to the introduced “complexity classes”, example quantifications of complexity could be given as follows:

- Flat documents and their retrieval based on the Vector Space Model could be measured by means of the VC dimension for linear classifiers, with $VC(H) = d + 1$ and d being the number of dimensions.
- Similarly, different VC bounds for graph structures exist. Such kind of bounds allow for deriving improved time bounds on shortest path queries [1], which points out the importance of quantifying information space complexity.
- Quantifying multimedia content and unlimited structural complexity remains an open issue for us. However, by considering retrieval as a topological problem, that is selecting a topological coherent region as answer for a query, topological measures like the Hausdorff-Besicovitch dimension may give possible estimates on information space complexity.

Obviously unsatisfactory is that the above classes and, consequently, their quantification, lack of a common scale or common measurement paradigm: given the short time available the working group could not identify a kind of language that underlies all forms of information needs. Perhaps such common basis does not exist at all (cf. the next subsection). Moreover, given accepted quantification measures for complex structures in IR, the question remains open as whether such a quantification yields general theoretical bounds on retrieval effectiveness and efficiency in practical applications.

4.3.3 Research Directions for Retrieval of Complex Structures

When humans use documents to communicate information, they use natural languages to express this information. Thus, while there is a variety of document flavors and domain-specific terminology, retrieval can still rely on the fundamental characteristics and regularities of human language. This explains why fairly generic techniques could be developed which can be applied successfully to a wide variety of document collections.

However, for complex structures this does not hold anymore: domain-specific multimedia content is frequently encoded in very specific ways which do not follow a generic language pattern (e.g., CAD models). But even textual information fragments in structured information don’t follow natural language patterns anymore, because the “information atoms” (such as a surname) are not connected in sentences, but live in isolated (database) fields.

Therefore, in the view of the working group participants, trying to create general-purpose retrieval systems for complex structures—in the same fashion in which we created general-

purpose document retrieval systems—seems to be a futile endeavor. But, although the complex structures itself are too heterogeneous to attempt a generalized retrieval approach, we observe that the information needs in various domains follow generic patterns. I.e., we can foster advances in this area by focusing on the process level instead of the system level. In this light, research can advance the field with three types of contributions:

1. by establishing well-defined processes for principled design of domain-specific complex structure retrieval systems,
2. by identifying and/or inventing reusable components (algorithms, patterns, frameworks) for the realization of such systems, and
3. by developing theoretical retrieval effectiveness and efficiency measures based on empirical complexity estimates.

A similar trend to adaptive retrieval strategies and customizable retrieval systems can be observed in the field of document retrieval (cf. Arjen’s talk).

4.3.4 Principled Design of Retrieval Processes on Complex Structures

A methodology guiding the design of retrieval processes on complex structures should cover several aspects. A core aspect is finding suitable representations for the complex structures at hand. If the representation is too elaborated, the search process will become cumbersome for users. On the other hand, if the model chosen is too compressed, users not finding what they need might not be able to refine their search appropriately. This applies not only to the representation of the information in the collection, but also to the representation of the information need (the query). In addition to supporting the model design, the methodology should also guide retrieval system developers in splitting the process into subtasks and selecting the right components for these subtasks. The working group discussed the following already existing components.

Faceted Search. Faceted search is successfully used to enable users expressing constraints on their search. It was noted that while facets are good as filter criteria, their usage for influencing the ranking of results is limited. One reason might be that the mechanics of score aggregation is not transparent for the average user.

Time/Space Representation. Time and space are fundamental concepts in understanding and structuring our experience and can be well expressed and presented to users. In case of the temporal dimension, a time line is the most natural visualization, while for spatial information the obvious visualization is a map. On the query side, constraints are expressed as time span and location-distance pair respectively.

Top-k and Skyline Queries. Top-k queries are suitable for structured data for which good relevance functions have been identified. The usage of Skyline queries is appropriate for low (2-5) dimensional data, when scoring functions for the individual data dimensions are available. However, they cannot be combined into a single relevance function.

Keyword queries on Structured Data. A lot of work has been done in the last years to extend the keyword search paradigm to structured data. Summarized, all these approaches view the structured data as graph and try to find small subgraphs (i.e., join paths) containing the query terms. Then they assess their relevance, and present the most relevant results to the user. While a lot of progress has been made, the retrieval quality level is still not very impressive [2]. Another direction in this area is the idea to infer the query intention in a multi-step retrieval process before assessing the query results (cf. Wolf’s talk).

User-defined Search Interfaces. In accordance with the core idea to support retrieval process design rather than to build generic retrieval systems for complex structures, frameworks for

retrieval have been built which allow the expert users (or system administrators) to specify their customized retrieval process (cf. Andreas' and Arjen's talks). These frameworks glue together individual retrieval components, and hence form a foundation for the transition of a retrieval process definition to a working retrieval system.

4.4 Sentiment and Opinion

Rapporteur: John A. Carroll

4.4.1 Tasks

Opinion mining can be viewed as 'natural' language-based task, as being something that a person might do in everyday life: finding out about what others think about a product, service, brand, organisation, or the views and actions of others. The information may come from a number of different types of document, such as online product reviews, newspaper editorials, blog posts, etc.

There are a number of opinion mining tasks, including:

- Subjectivity classification – whether a sentence, paragraph or whole document is expressing some sort of opinion
- Polarity (or sentiment) classification – whether an opinion being expressed is broadly positive or negative, or positive / neutral (i.e. non-opinionated) / negative; other more fine-grained classification schemes have also been used
- Relevance – whether a particular document is relevant to a given topic of interest
- Authority – the degree of influence of the opinions expressed in a document on its readers
- Opinion holder and target identification – who is expressing an opinion, and about what
- Sentiment towards features of a product or aspects of a topic – for example whether the picture quality, price, and battery life of a camera are good or bad; or whether government policy on taxation is good or bad
- Detecting key messages related to a brand or marketing campaign – reflecting the impact that it has had
- Classification of documents along other affective dimensions – such as anger, frustration, or satisfaction
- Other opinion-related tasks – for example rating the perceived 'green-ness' of politicians, or the proportion of people in favour of building new nuclear plants

In the past many of these tasks were carried out by human analysts, originally working from printed media; with the increasing amounts of online text and capabilities of language processing systems, automatic opinion mining has become a highly active area of basic and applied research.

4.4.2 Barriers to Progress

Data Quality. Most research in sentiment classification has used data consisting of reviews of products or services scraped from review websites, for example IMDb (the Internet Movie Database), Amazon, or epinions. When review authors give 'star' ratings to accompany their reviews, this can result in large amounts of data annotated at the document level.

In contrast, there is very much less text available that is annotated for sentiment at more detailed levels (sentence or phrase), or annotated to support other opinion mining tasks.

Notable exceptions include the MPQA Opinion Corpus and the datasets produced for the NTCIR MOAT workshops. However, it is difficult to satisfactorily circumscribe some opinion mining tasks: for example in opinion holder and target identification, should holder and target phrases include post-modifiers, and if so what types of post-modification? Without well-defined annotation guidelines, inter-annotator agreement will be poor leading to a low upper bound on system performance.

Another approach to obtaining annotated data is crowd-sourcing, for example via the Amazon Mechanical Turk. The quality of annotations obtained this way can be questionable, but may be assured by framing the task carefully, for instance requiring that a quote from the document be pasted into a text box to justify the sentiment annotation being entered.

Evaluation. Funding is hard to obtain for long-term evaluation campaigns or shared tasks. There are only two of these worldwide which contain an opinion mining component: TREC (blog track) and NTCIR (multilingual opinion analysis task), organised by NIST and NII respectively. Other evaluation exercise series that have had opinion-related components include CLEF, SemEval, and i2b2. These exercises often rely for annotation on ad-hoc groups of volunteers – usually the participants themselves. Short timescales and lack of funding make it difficult to generate significant amounts of high quality data, limiting the advances that can be made.

An important methodological issue (not restricted to this research area) is reproducibility of experiments. Datasets and systems are not always publicly available, and algorithms and parameter settings are often not described in sufficient detail to be able to reproduce results. This may make it difficult or even impossible for subsequent researchers to conduct appropriate comparative evaluations, hindering progress.

4.4.3 Research Priorities

Cross-domain. Accuracy of sentiment classification can be severely degraded if a system trained on data from one domain is applied to a distinct domain. However, there is often insufficient annotated data in a new domain to adequately train a supervised machine learning algorithm. Domain adaptation techniques, such as structural correspondence learning (e.g. work by Blitzer) and related approaches are a promising way of tackling this important issue. These techniques use annotated data in a source domain in conjunction with unannotated data in the target domain (and possibly also a further small amount of annotated data in the target domain).

Cross-language. Compared with most languages, there is a large amount of data in English that could be used for developing opinion mining systems. Cross-language systems take advantage of this by mapping information from a source language (e.g. English) to the target language. The mapping can be done with techniques similar to those used for domain adaptation; however this is a difficult problem since different cultures may express opinions using different types of language.

Novel sources of data. It is relatively easy to obtain suitable annotated data for sentiment classification of product reviews (e.g. from review sites' star ratings), but it is in general difficult for other opinion mining tasks. Finding good sources of annotations avoids the need for costly annotation efforts. Examples of such sources are: metadata for authority (for example '10 out of 15 people found this useful'), and summaries of key points abstracted from longer reviews in special interest magazines (such as about cars or yachts).

Emergent features. Given some knowledge of a type of product or service it is possible to come up with a list of features that reviewers might have an opinion on, and develop a system

to find opinions on these; however, it is also important to be able to capture ‘emergent’ features that people have an opinion on but which hadn’t been anticipated.

Presentation of results. Traditionally, human analysts in media monitoring companies write reports for clients which contain a range of quantitative and qualitative information presented as text, tables and graphics. An automatic system could not currently produce such reports. The question then is how to present the results of opinion mining. Some possibilities are: overall aggregate sentiment, a selection of typical documents, a set of features and the sentiment ascribed to each, opinion associated with sub-topics in the domain, or opinion-driven summarisation. A further question is to do with explanatory level: why does an individual perceive something as being good or bad? This relates to sentiment-specific aspects of entailment.

More detailed analysis. Media monitoring companies have traditionally used a 5-category sentiment classification scheme (strongly / weakly negative, neutral, and strongly / weakly positive), rather than the 2- or 3-category scheme usually adopted in computational work. More fine-grained analysis requires deeper representations than the typical ‘bag of words’ and deeper processing than supervised classification over these representations. Such analysis might involve (at least) parsing, integration of contextual valence, and principled treatments of negation and hedging.

4.4.4 Summary

Opinion mining has many different facets and presents numerous difficulties, as outlined above. A lot of progress has been made, but as in most areas of document mining, there are still many outstanding research challenges.

References

- 1 Ittai Abraham, Daniel Delling, Amos Fiat, Andrew V. Goldberg, and Renato F. Werneck, *VC-Dimension and Shortest Path Algorithms*, In Proceedings of ICALP 2011.
- 2 J. Coffman and A. C. Weaver, *A framework for evaluating database keyword search strategies*, In Proceedings of CIKM 2010.
- 3 V. Vapnik and A. Chervonenkis. *On the uniform convergence of relative frequencies of events to their probabilities*. Theory of Probability and its Applications, 16(2):264-280, 1971.

5 A Theme: Towards More Open Search In Europe

The European Union is investing more than 5 billion euros in Galileo, a global navigation system similar to the US-owned GPS. One of the motivations for this investment is the potential impact of a GPS shutdown (perhaps in time of crisis or a period of intense trade disputes or similar). Far-fetched, perhaps, but then the disappearance of hundreds of billions of euros into a black hole of financial chaos appeared far-fetched until recently.

An obvious corollary question is this:

What would be the economic cost to the EU of a shutdown of the search engines provided by Google, Yahoo and Microsoft?

As the World Wide Web has become a key repository of human knowledge, interaction and commerce, so the ability to discover relevant pages has become a core function of our economies. It is now routine to describe 21st century industry as a *knowledge economy*, but what price knowledge if it is impossible to find? How many of the web addresses that you use (those ‘uniform resource locators’ that start with `http://`) can you remember?

Let's imagine a sudden restriction of service from the big three search providers (effectively now a 'big two' after the deal between Yahoo and Microsoft). Technologically speaking removing service availability based on region can be done trivially – witness the current mechanisms to control access to BBC programmes abroad, for example.

The economic consequences for the EU would be immediate and wide-reaching. Our elected representatives and state bureaucrats would be tasked with finding a replacement – but their first port of call would be the search box in their web browser...

Secondly, there is increasing concern that detailed individual and private data is becoming more and more centralised and open to abuse. Let's imagine that we suffer a number of scandalous cases at some future point, and the legal framework changes to the degree that the current models no longer work. At present the EU would be powerless to implement the requisite technological changes.

We are not the first to recognise this achilles heal in European independence. The Quaero initiative announced by Jacques Chirac and Gerhard Schröder in 2005 aimed to fill the gap, but by 2007 *The Economist* reported that the project had effectively been scrapped (for various reasons – but the huge cost of the compute infrastructure and bandwidth consumption required must be a large factor). Several results have survived: multimedia retrieval projects in Germany and the Exalead engine in France, but the dream of triggering construction of a pan-European search engine has remained a dream.

What to do?

Resurrecting the vision of Quaero and a full-blown European competitor for Google and Microsoft/Yahoo is a matter for forces more powerful than those we represent or can hope to influence. However, all is not lost, partly because of the changing landscape of the web, of utility computing, and of social interaction in the networked age. If Europe can find the resources to fund a set of incremental improvements in our search R&D then we can take a major step towards greater openness and independence from the dominant corporate US infrastructures.

We propose these measures:

1. Support the use of cloud computing to develop national search service platforms within existing information retrieval research computing centers. Long-term commitment is more important than a dramatic expansion of funding; the current situation is short-termist and discourages robust search provision.
2. Create new marketplaces for value-added services on top of the cloud infrastructure. In particular the need for analysis of social media is becoming very widespread (for example: customer relations used to be about sitting next to the phone or the cash till; now its about reading 250,000 Twitter posts per week).
3. Develop aggregation and syndication models of distributed clustering and indexing to allow the composition of international and cross-domain search services, drawing on the above provisions. (For example recent work on peer-to-peer indexing and clustering in web servers like Apache.)

The cost of an alternative to Google is probably too large a mouthful to swallow in one go, but a smaller set of chunks can bring us a solution in the short to medium term.

5.1 Discussion

- issues with 3. (P2P): spam; speed (necessitates regional/national aggregation)
- grey literature, intranets: organisations can create custom search views that incorporate their own private data

- richer model of privacy, the ability to buy privacy (currently we sell privacy – we take free services in return for giving up some privacy – but have no alternative if we want to buy privacy)
- what about the energy efficiency of running the big 4 web companies (Google, M\$/Yahoo, Facebook, Twitter) in a single country?

6 Recap of the Proposal

Document mining is the process of deriving high-quality information from large collections of documents like news feeds, databases, or the Web. Document mining tasks include cluster analysis, classification, generation of taxonomies, information extraction, trend identification, sentiment analysis, and the like. Although some of these tasks have a long research history, it is clear that the potential of document mining is far from full realisation.

Part of the problem is that relevant document mining techniques are often applied in an isolated manner, addressing – from a user perspective – only a part of a task. E.g., an intelligent cluster analysis needs adequate document models (information retrieval) that are combined with sensible merging algorithms (unsupervised learning), complemented by an intuitive labelling (information extraction, natural language processing). The deficit that we observe may also be understood as a lack of application- and user-orientation in research.

In this seminar we want analyze the untapped potential. To this end we bring together researchers from the main areas of document mining to present their view, to understand where and how latest disciplinary achievements can be combined, and to develop a more integrative view on document mining.

The seminar is especially timely as in recent years the field has grown by a large amount, and this has lead to increasing specialisation of the research community. Now is the time to bring a sample of the leading teams back together and look at the problem from a multidisciplinary point of view

The seminar will focus on the following aspects of document mining:

1. What are the relevant document mining tasks? The expectations and the potential for document mining changed significantly over time. Influential in this connection is the discovery of the enormous contributions of users to the Web (among others in the form of blogs, comments, reviews) as highly valuable information source.
2. Cluster analysis is a key technology in document mining and involves several issues on its own, which are detailed below. A major deal of cluster analysis research has been spent to merging principles and algorithms; today, and especially in document mining, the research focus is on tailored document models, user integration, topic identification and cluster labelling, on the combination with retrieval technology (e.g. as result set clustering), or on support technology for supervised classification. Moreover, theoretical foundations of cluster analysis performance in document mining as well as commonly accepted optimality measures are open research questions.
3. While cluster analysis can be mainly considered as unsupervised, several advanced document mining tasks combine unsupervised with supervised text classification. Especially non-topical classification tasks attracted interest in this connection, such as genre classification, sentiment analysis, or authorship grouping.
4. Document mining poses various challenges from a machine learning perspective. An important constraint is the lack of sufficient amounts of labelled data. This situation will become even more unbalanced in the future, and current research (domain transfer

- learning, transductive learning) aims at the development of technology to exploit the huge amount of unlabelled data to improve supervised classification.
5. Robustness and efficiency of document mining technology is a key issue from the user perspective and for future applications. Both may be achieved by the combination of algorithms (ensemble clustering) or the combination of different retrieval models; the respective research is still at its infancy.
 6. The use of NLP in document mining is a success factor of increasing importance for document mining. Among others this field contributes technology for document modelling, style quantification, document segmentation, topic identification, and various information extraction and semantic annotation tasks.
 7. The assessment of information quality and credibility will have a large impact on future document mining solutions. It can tackle information need issues and performance problems in our information-flooded society at the same time. However, there are various open research question related to the measurability of information quality.
 8. Authorship and writing style modelling is still coming of age; this area forms the heart for high-level document mining tasks such as plagiarism analysis, authorship attribution, and information quality assessment.
 9. Future technology for entity resolution and fact or relation extraction. Current approaches are limited to closed environments where the target entities and relations are known in advance. In practice, however, this assumption is often violated or the number of entities and relations is so large that automatic methods are needed. The Open Information Extraction paradigm, coined by Banko and Etzioni, addresses these issues. Its goal is to extract a diverse set of relational tuples from text without any relation-specific input such as hand-labeled examples or hand-crafted lexico-syntactic patterns.
 10. Interface design and visualization are very important for effective user access to the output of the document mining process. Moreover, interactive document mining approaches like e.g. scatter-gather clustering pose new challenges for both the interface and the backend.
 11. Finally, evaluation is essential for developing any kind of data mining method. So far, mainly system-oriented evaluation approaches have been used, where the data mining output is compared to some ‘gold standard’. There is a lack of user-oriented evaluations (e.g. observing users browsing a cluster hierarchy), that also take into account the tasks the users want to perform – e.g. using Borlund’s concept of simulated work tasks.

The general idea is to collect the state of the art in document mining research, and to define a research agenda for further work in this area. For this purpose, we want to bring together experts from the different research areas listed above. Each of the participants first should present her/his view on particular document mining challenges by highlighting latest results and naming crucial research issues. Based on these contributions, we will aim at developing an integrated view on the problem of document mining, identify open research problems and then point out steps towards resolving these problems. Altogether we would like to come up with a framework that associates document mining tasks with the scientific and technological elements of their adequate solution.

7 Acknowledgements

Our thanks to the staff of Schloss Dagstuhl for excellent organisation and facilities.

Participants

- Leif Azzopardi
University of Glasgow, GB
- Ted Briscoe
University of Cambridge, GB
- Steven Burrows
Bauhaus-Universität Weimar, DE
- John A. Carroll
Univ. of Sussex – Brighton, GB
- Massimiliano Ciaramita
Google Switzerland – Zürich, CH
- Hamish Cunningham
Sheffield University, GB
- Arjen P. de Vries
CWI – Amsterdam, NL
- Norbert Fuhr
Universität Duisburg-Essen, DE
- Tim Gollub
Bauhaus-Universität Weimar, DE
- Thomas Gottron
Universität Koblenz-Landau, DE
- Michael Granitzer
Know-Center Graz, AT
- Andreas Henrich
Universität Bamberg, DE
- Gerhard Heyer
Universität Leipzig, DE
- Dennis Hoppe
Bauhaus-Universität Weimar, DE
- Melikka Khosh Niat
Universität Duisburg-Essen, DE
- Marc Lechtenfeld
Universität Duisburg-Essen, DE
- Alexander Löser
TU Berlin, DE
- Peter Prettenhofer
TU Graz, AT
- Andreas Rauber
TU Wien, AT
- Harald Reiterer
Universität Konstanz, DE
- Stefan M. Rieger
The Open University – Milton Keynes, GB
- Hinrich Schütze
Universität Stuttgart, DE
- Wolf Siberski
Leibniz Univ. Hannover, DE
- Benno Stein
Bauhaus-Universität Weimar, DE



Artificial Immune Systems

Edited by

Emma Hart¹, Thomas Jansen², and Jon Timmis³

1 Napier University – Edinburgh, GB, e.hart@napier.ac.uk

2 University College Cork, IE, t.jansen@cs.ucc.ie

3 University of York, GB, jtimmis@cs.york.ac.uk

Abstract

This report documents the program and the outcomes of the Dagstuhl Seminar 11172 “Artificial Immune Systems”. The purpose of the seminar was to bring together researchers from the areas of immune-inspired computing, theoretical computer science, randomised search heuristics, engineering, swarm intelligence and computational immunology in a highly interdisciplinary seminar to discuss two main issues: first, how to best develop a more rigorous theoretical framework for algorithms inspired by the immune system and second, to discuss suitable application areas for immune-inspired systems and how best to exploit the properties of those algorithms.

Seminar 26.–29. April, 2011 – www.dagstuhl.de/11172

1998 ACM Subject Classification I.2.8 Problem Solving, Control Methods, and Search; Heuristic methods.

Keywords and phrases Artificial Immune Systems, Randomised Search Heuristics


Digital Object Identifier 10.4230/DagRep.1.4.100

1 Executive Summary

Emma Hart

Thomas Jansen

Jon Timmis

License  Creative Commons BY-NC-ND 3.0 Unported license
© Emma Hart, Thomas Jansen, Jon Timmis

Artificial immune systems (AISs) are inspired by biological immune systems and mimic these by means of computer simulations. They are seen with interest from immunologists as well as engineers. Immunologists hope to gain a deeper understanding of the mechanisms at work in biological immune systems. Engineers hope that these nature-inspired systems prove useful in very difficult computational tasks, ranging from applications in intrusion-detection systems to general optimization. Moreover, computer scientists identified artificial immune systems as another example of a nature-inspired randomized search heuristic (like evolutionary algorithms, ant colony optimization, particle swarm optimization, simulated annealing, and others) and aim at understanding their potential and limitations. While the relatively new field has its successful applications and much potential its theoretical foundation is still in its infancy. Currently there are several not well connected strands within AIS theory, not even a general agreement on what the central open problems are, and only a weak connection between AIS theory and AIS applications. The main goals of the proposed seminar include bringing together computer scientists and engineers to strengthen the connections within AIS theory, connections to other researchers working on the theory of randomized search heuristics, and to improve connectivity between AIS theory and applications.



Except where otherwise noted, content of this report is licensed under a Creative Commons BY-NC-ND 3.0 Unported license
Artificial Immune Systems, *Dagstuhl Reports*, Vol. 1, Issue 4, pp. 100–111
Editors: Emma Hart, Thomas Jansen, and Jon Timmis



DAGSTUHL REPORTS
Dagstuhl Reports
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

Biological immune systems show great resilience in harsh environments and demonstrate the ability to cope with large amounts of sensory data as well as the unpredictability of the natural world. Indeed, a great deal of attention is now being paid to these aspects of the immune system by the wider computing research community.

Given the practical success of AIS, there is a serious lack of theoretical work in the area. Many AIS algorithms are based purely on clonal selection mechanisms, without any interaction between the different members of the cell populations. The dynamics of cell populations in the immune system have been modeled extensively using nonlinear dynamical systems. At present, however, there is no centrally agreed approach on how to tackle important theoretical issues in AIS. All too often theory is undertaken without the due attention to the practical implications. For theory to have a serious impact, collaboration between theoreticians and engineers is needed to identify key engineering issues, relevant theoretical issues and crucially how the theory can help support the engineering process. While starting point of the seminar and its driving force are deficits in the theoretical foundation of AIS its main goals are clearly beyond theory. At the heart of the seminar's motivation is the conviction that there is nothing more practical than a good theory.

The seminar took place from April 26th to April 29th 2011. It started with a series of talks aimed at providing a suitable level of introduction to the main areas of discussion to provide a levelling ground for all participants. The format of the seminar was then a series of short presentations by researchers on topics that ranged from swarm robotics to immunology and theoretical frameworks for algorithm analysis. These were then followed by a series of *breakout* group sessions which focussed discussion on the issues raised by the speakers with results from those discussions being reported back to the main group at regular intervals. Towards the end of the week, a convergence into four key topics emerges: (1) The principled development of bio-inspired algorithms and how the translation from computational models into usable algorithms is managed, (2) the relationship between evolution and immunity and how it might be possible to evolve an artificial immune system in complex engineering problems, specifically swarm robotic systems, (3) the development of a definitive clonal selection algorithm with appropriate theoretical analysis and (4) the development of novel immune algorithms and the use of models from computational immunology for both the understanding of immunological processes and the development of new algorithms. These four topics are to be taken forward as journal papers by participants from the seminar.

As a result of the seminar there will be a special issue published in *Natural Computing* a leading journal in the area that will not only publish papers outlined above, but provide a roadmap for the future direction of AIS and serve as, it is hoped, an authoritative guide to the area of artificial immune systems.

2 Table of Contents

Executive Summary

Emma Hart, Thomas Jansen, Jon Timmis 100

Overview of Talks

Training a network of mobile neurons
Bruno Apoloni 103

Immune inspired approaches to ab initio modelling
George M. Coghill 103

Theory of Randomized Search Heuristics
Benjamin Doerr 104

Making Affective Effective: Emotion Classification and AIS
Julie Greensmith 104

Applications of Immune Inspired Computing
Emma Hart 105

A Crash Course in Immunology and AIS
Emma Hart 105

Swarm-based Modeling and Visualization of the Immune System ... and more
Christian Jacob 106

AIS Theory
Thomas Jansen 106

Runtime Analysis of Clonal Selection Algorithms
Per Kristian Lehre 107

Real and artificial immune systems: oil and water or milk and cookies?
Chris McEwan 107

Less Bio, More Inspired
Robert Oates 108

The use of AIS techniques to protect an Artificial Hormone System
Mathias Pacher 108

Negative Selection Algorithms: Past, Present, Perspectives
Johannes Textor 108

On Fault Tolerance and Scalability of Swarm Robotic Systems
Alan FT Winfield 109

Run Time Analysis of Artificial Immune Systems
Christine Zarges 109

Working Groups

A principled approach to deriving immune inspired engineering principles
Marc Read, Chris McEwan, Emma Hart, Ed Clark, Julie Greensmith, Uwe Aickelin 110

Participants 111

3 Overview of Talks

3.1 Training a network of mobile neurons

Bruno Apoloni (Università di Milano, IT)

License © © ⊖ Creative Commons BY-NC-ND 3.0 Unported license
© Bruno Apoloni

Joint work of Apoloni, Bruno; Simone, Bassis; Lorenzo, Valerio

Main reference Apoloni, Bruno; Simone, Bassis; Lorenzo, Valerio, "Training a network of mobile neurons," Proc. IJCNN 2011

We introduce a new paradigm of neural networks where neurons autonomously search for the best reciprocal position in a topological space so as to exchange information more profitably. The idea that elementary processors move within a network to get a proper position is borne out by biological neurons in brain morphogenesis. The basic rule we state for this dynamics is that a neuron is attracted by the mates which are most informative and repelled by ones which are most similar to it. By embedding this rule into an Newtonian dynamics, we obtain a network which autonomously organizes its layout.

Thanks to this further adaptation, the network proves to be robustly trainable through an extended version of the backpropagation algorithm even in the case of deep architectures.

We test this network on two classic benchmarks and thereby get many insights on how the network behaves, and when and why it succeeds.

3.2 Immune inspired approaches to ab initio modelling

George M. Coghill (University of Aberdeen, GB)

License © © ⊖ Creative Commons BY-NC-ND 3.0 Unported license
© George M. Coghill

Joint work of Coghill, George M.; Pang, Wei

Main reference Pang, W. & Coghill, GM. (2011), "An immune-inspired approach to qualitative system identification of biological pathways," Natural Computing, vol 10, no. 1, pp. 189–207.

URL <http://dx.doi.org/10.1007/s11047-010-9212-2>

Model-based systems and qualitative reasoning (MBS&QR) provides a means of reasoning about the structure and behaviour of dynamic systems in situations where information about the system of interest is sparse or incomplete. In domains such as systems biology the data available are mixed: some are sparse, some are detailed and some are qualitative. In such circumstances MBS&QR may enable one to simulate, construct or modify imprecise models of metabolic systems.


We have developed a number of MBS&QR tools to facilitate this: a fuzzy qualitative reasoning system to provide abstract simulations, diagnostic machinery and Artificial Immune System based model learners. For this latter the results of the learning are very promising for ab initio system identification.

A version utilizing CLONALG generates results comparable to deterministic model learning for small models, but which comes into its own for larger systems.

Comparative studies with an improved version of optAINet reveal improved identification speeds for larger systems

3.3 Theory of Randomized Search Heuristics

Benjamin Doerr (MPI für Informatik – Saarbrücken, DE)

License  Creative Commons BY-NC-ND 3.0 Unported license
© Benjamin Doerr

In this survey talk, I will give a high-level summary of what happened in the area of theoretical analyses of randomized search heuristics. After the talk, I'm hoping for interesting discussions on whether this can serve as a roadmap for the developing theory of artificial immune systems or what should be done differently.

By theory, I shall adopt the strict view of proving precise statements by mathematical means. I shall then highlight three research directions that were followed in the theory of randomized search heuristics.

Artificial example problems: These did a good job in refuting common misbeliefs, for example, that unimodal functions are always easy to optimize via hill-climbers.


Run-time analysis. Inspired by classical randomized algorithms theory, we may also try to bound the run-time of a randomized search heuristic. Usually, we do not regard the actual run-time, but count the number of fitness evaluations.

This reflects the common assumption that this is the most costly part in many randomized search heuristic applications.

Black-box complexity: Since classic algorithms theory greatly profited from a powerful complexity theory, a similar methodology might be helpful for randomized search heuristics as well. If again we take the number of search point evaluations as complexity measure, this leads to the notion of (unrestricted) black-box complexity. It is currently discussed to what extent this notion yields useful insight on the problem difficulty for randomized search heuristics. Most likely, the class of all black-box optimization algorithms should be restricted in a way that excludes overly powerful, but unrealistic heuristics.

3.4 Making Affective Effective: Emotion Classification and AIS

Julie Greensmith (University of Nottingham, GB)

License  Creative Commons BY-NC-ND 3.0 Unported license
© Julie Greensmith

URL <http://www.ima.ac.uk/greensmith>

Affective computing is an emerging multidisciplinary field within computer science with the aim of incorporating of emotions into computational systems.

The nature of this incorporation can occur via the use of emotion data from users of such systems. Equally it can represent the display of emotions by computational systems to enhance the experience of the user through the generation of an emotive dialogue.


There are a number of challenges in working with affective systems centered on using emotions. Emotions are dynamic, they change over time depending on the internal and external milieu of the individual. Emotions are personal, with individuals portraying different emotions in a number of different physiological and social ways. Emotions are subjective, as it is difficult to find a universal taxonomy or strict definition for any individual emotion. Specific identifiable patterns or physiological signatures of emotion simply do not exist between individuals.

This makes it an interesting problem in terms of the application of computational intelligence techniques. However, little research exists in the area of the application of

intelligent classification techniques to the ascertainment of emotions from physiological data. In this talk I presented the wearable biosensor technology used to collect physiological data and propose an approach for the modelling and classification of this data using ensemble based artificial immune system classification methods. In addition I proposed and gained feedback on a framework combining AIS algorithms to process multiple levels and types of emotional response.

3.5 Applications of Immune Inspired Computing

Emma Hart (Edinburgh Napier University, GB)

License  Creative Commons BY-NC-ND 3.0 Unported license
© Emma Hart

Joint work of Hart, Emma; Timmis, Jon

Main reference Hart,E and Timmis,J., “Applications of AIS: The Past, the Present and the Future,” *Applied Soft Computing* 8(1), 2008, pp. 191–201

URL <http://dx.doi.org/10.1016/j.asoc.2006.12.004>

The field of Artificial Immune Systems (AIS) has seen a number of successful algorithms developed over the past 20 years. AIS algorithms can be considered as either evolutionary or swarm like and they are inspired by a variety of immunological processes. Recently there has been a drive to provide theoretical underpinnings of some of these algorithms. Despite these successes, in this talk we highlight issues that might potentially hinder further development of the field. We explicitly separate immune properties and immune functionality and use this classification to highlight potential of engineered systems that can exploit either immune properties or immune function. We discuss ad-hoc sensor networks and swarm robotic systems that naturally map to suitable immune functions and properties. We highlight issues of translating immune models into engineered immune-inspired solutions. We argue that functionality driven design will lead to systems that don't necessarily replicate biological mechanisms but replicate biological functionality.

3.6 A Crash Course in Immunology and AIS

Emma Hart (Edinburgh Napier University, GB)


License  Creative Commons BY-NC-ND 3.0 Unported license
© Emma Hart

Joint work of Hart, Emma; Jon Timmis

This talk provides a high level introduction to immunology and to the algorithms and application areas that have been developed from looking to the immune system for inspiration. It introduces immune mechanisms such as negative selection, clonal selection and danger theory. These are related to computational algorithms that have been developed to address applications in optimisation, anomaly detection and classification.

3.7 Swarm-based Modeling and Visualization of the Immune System ... and more

Christian Jacob (University of Calgary, CA)

License  Creative Commons BY-NC-ND 3.0 Unported license
© Christian Jacob

Joint work of Jacob, Christian; von Mammen, Sebastian; Davison, Timothy; Sarraf, Abbas; Sarpe, Vladimir; Esmaili, Afshin; Phillips, David; Yazdanbod, Iman

The Lindsay Virtual Human (LINDSAY) project creates a 3-dimensional, interactive computer model of male and female anatomy and physiology for medical education. The software developed in the LINDSAY project provides an exploration tool for medical students that complements their experiences and learning in the classroom, with simulators, and with patients.

Lindsay Presenter (LPresenter) is our first prototype of a presentation tool for interactive, 3-dimensional anatomical contents, to be used by instructors in medical schools. LPresenter can also be used by students to review contents from the lectures, prepare for exams, use quiz-type inquiry, or help with general as well as specific inquiry of anatomical structures. LPresenter has a built-in anatomy database and searchable atlas, which provides easy access to anatomical contents.

Lindsay Composer (LComposer) provides access to simulations of physiological processes. LComposer is the software tool that integrates physiology contents into medical teaching and learning resources as a computer-based modeling and exploration tool for human physiology. Similar to LPresenter, LComposer allows to program and compose physiological scenarios, which would illustrate key concepts of human physiology related to medical education. LComposer also incorporates a graphical programming environment, so that simulations and computer models can be assembled without any prior programming experience.

I presented first results from our LINDSAY Virtual Human project and demonstrated our first steps towards integrating simulations of immune system processes into multi-scale physiology simulations.

3.8 AIS Theory

Thomas Jansen (University College Cork, IE)

License  Creative Commons BY-NC-ND 3.0 Unported license
© Thomas Jansen

With help by and thanks to





- Uwe Aickelin, University of Nottingham,
- Ed Clark, University of York,
- Robert Oates, University of Nottingham,
- Thomas Stibor, TU München,
- Johannes Textor, Universität Lübeck,
- Christine Zarges, TU Dortmund.

Aim of the brief overview talk is to provide a very basic orientation in the field of artificial immune system and relevant theoretical approaches. As four main topics clonal selection, negative selection, dendritic cells and immune networks are identified. For each of these we identify a biological basis, prominent systems or algorithms, and theoretical approaches and

provide pointers to literature. As key issues in AIS theory we briefly discuss the aspects of equivalence to other known approaches, affinity and complexity.

3.9 Runtime Analysis of Clonal Selection Algorithms

Per Kristian Lehre (Technical University of Denmark, DK)

License     Creative Commons BY-NC-ND 3.0 Unported license
© Per Kristian Lehre





The recent progress in runtime analysis of evolutionary algorithms (EAs) has been facilitated by the introduction of appropriate analytical techniques. However, many of these techniques are primarily suited for the analysis of EAs with a parent population size of one, and not the population-based EAs that are often used by practitioners.

The first part of the talk gives a brief introduction to some analytical techniques that have recently become available for EAs with larger population sizes (Lehre PPSN2010 & GECCO2011). In particular, we introduce a variant of the well-known artificial fitness level technique which can be used to derive upper bounds the expected runtime of non-elitist EAs with populations.

In the second part of the talk, we discuss whether these techniques can be helpful for researchers in the artificial immune systems community, in particular for the study of clonal selection algorithms.

3.10 Real and artificial immune systems: oil and water or milk and cookies?

Chris McEwan (Edinburgh Napier University, GB)

License     Creative Commons BY-NC-ND 3.0 Unported license
© Chris McEwan


The seminal work in artificial immune systems set out to establish a productive interface between computer science and immunology. As the field gained popularity, the focus of research shifted towards "novel approaches" to applied problems in optimisation and artificial intelligence. With this shift, the once shared goal of hypothesising how the immune system might realise such functionality was compromised, and with it, any opportunity for AIS to establish itself as an independent body of work that contributes novel ideas to the immunological and engineering domains.

In this talk, I will address better aligning the real and artificial immune systems. This approach centers on the role of theory in providing mechanistic explanations behind such systems. I will argue that AIS should always be able to provide a biological interpretation, regardless of validity or plausibility; otherwise, they are only trivially "immune inspired". At the same time, AIS should be communicable with minimal references to biological nomenclature; otherwise, it is difficult to assert novelty and worth as computing artifacts.

Only once both desiderata are met can one claim to have achieved "immune-inspired computing". To meet both at the same time requires theoretical support for moving between the two mindsets. I contend that here lies the current and future contribution of AIS as a body of work.

3.11 Less Bio, More Inspired


Robert Oates (University of Nottingham, GB)

License  Creative Commons BY-NC-ND 3.0 Unported license
© Robert Oates

This talk represents a deliberately hyperbolic argument against the utility of the field of AIS and bio-inspired computing in general, in order to stimulate debate about the field's short-comings and, hopefully, inspire a well-structured counter-argument. It highlights issues surrounding the fidelity of the underlying biological models to the systems they purport to represent and questions the "cell-by-cell" approach to developing immune-inspired algorithms. Far from concluding, the talk ends with a series of open questions to act as jumping off points for a deeper audience discussion.

3.12 The use of AIS techniques to protect an Artificial Hormone System

Mathias Pacher (Universität Frankfurt am Main, DE)

License  Creative Commons BY-NC-ND 3.0 Unported license
© Mathias Pacher


Joint work of Pacher, Mathias; Brinkschulte, Uwe

We present an Artificial Hormone System (AHS) which is able to allocate tasks on resources of a distributed system. The AHS holds "organic" properties like self-configuration, self-optimization, and self-healing. We proved that the AHS works perfectly well under normal circumstances. However, in case of hardware failures or malicious attacks the effects on the AHS may be fatal: Its real-time bounds may be violated or there may be task loss or system overload which leads to damages of the resources of the distributed system.

We try to use or adapt AIS algorithms to counteract the effects arising from failures and attacks. We show a first approach demonstrating that AIS algorithms may help protect the AHS.

3.13 Negative Selection Algorithms: Past, Present, Perspectives

Johannes Textor (Universität Lübeck, DE)

License  Creative Commons BY-NC-ND 3.0 Unported license
© Johannes Textor

Joint work of Textor, Johannes; Elberfeld, Michael; Liśkiewicz, Maciej

Main reference M. Liśkiewicz, J. Textor, "Negative selection algorithms without generating detectors," Proc. Genetic and Evolutionary Computation Conference (GECCO'10), pp. 1047–1054, ACM, 2010.

URL <http://dx.doi.org/10.1145/1830483.1830673>

Negative selection algorithms are immune-inspired binary classifiers that are trained on only negative examples. Though initially a considered promising approach, there was so far no success in applying negative selection algorithms to real-world machine learning problems. One of the major obstacles was the computational demand, which was generally exponential in the size of the input data.

By casting negative selection algorithms in the formal framework of algorithmic learning theory, we showed that the while efficient implementations of the scheme are indeed unlikely to exist, one can efficiently *simulate* the scheme, e.g. by constructing a finite automaton whose accepted language is equal to the set of all strings that the algorithm would assign a positive label to.

We give a brief overview of how this approach provided new insight into the computational complexity of negative selection algorithm - based classification for some important classes of patterns that were used in the artificial immune systems literature. For some pattern classes, including the prominent *r*-chunk and *r*-contiguous patterns, we provide polynomial time solutions while previously, there were only exponential time implementations available.

Moreover, we discuss how negative selection algorithms are becoming increasingly important tools in theoretical immunology, where they function as a formal model of the negative selection process itself.

References

- 1 M. Liśkiewicz and J. Textor. Negative selection algorithms without generating detectors. In *Proceedings of Genetic and Evolutionary Computation Conference (GECCO'10)*, pages 1047–1054. ACM, 2010.
- 2 M. Elberfeld and J. Textor. Negative selection algorithms on strings with efficient training and linear-time classification. *Theoretical Computer Science*, 412:534–542, 2011.

3.14 On Fault Tolerance and Scalability of Swarm Robotic Systems

Alan FT Winfield (University of the West of England - Bristol, GB)

License © © ⊖ Creative Commons BY-NC-ND 3.0 Unported license
© Alan FT Winfield

Joint work of Bjercknes, Jan Dyre; Winfield, Alan FT

Main reference Bjercknes JD and Winfield AFT, “On Fault-tolerance and Scalability of Swarm Robotic Systems,” Proc. Distributed Autonomous Robotic Systems (DARS 2010), Lausanne, November 2010.

There is a common assumption that swarm robotic systems are robust and scalable by default. This talk will present an analysis based on both reliability modelling and experimental trials of a case study swarm performing team work, in which failures are deliberately induced. The case study has been carefully chosen to represent a swarm task in which the overall desired system behaviour is an emergent property of the interactions between robots, in order that we can assess the fault tolerance of a self-organising system. Our findings show that in the presence of worst-case partially failed robots the overall system reliability quickly falls with increasing swarm size. We conclude that future large scale swarm systems will need a new approach to achieving high levels of fault tolerance.

3.15 Run Time Analysis of Artificial Immune Systems

Christine Zarges (TU Dortmund, DE)

License © © ⊖ Creative Commons BY-NC-ND 3.0 Unported license
© Christine Zarges

In this talk, artificial immune systems, in particular clonal selection algorithms used for optimization, are considered from the perspective of general randomized search heuristics.


An overview on recent theoretical results on the run time of simple algorithms is presented.

In the first part, we discuss several mutation operators that are used in practical algorithms and point out benefits and drawbacks of such mechanisms. In the second part, the use of the concept of static pure aging is investigated and compared to a similar mechanism from the field of evolutionary computation. We close by pointing out interesting and important aspects for future work.

4 Working Groups

4.1 A principled approach to deriving immune inspired engineering principles

Marc Read, Chris McEwan, Emma Hart, Ed Clark, Julie Greensmith, Uwe Aickelin

License  Creative Commons BY-NC-ND 3.0 Unported license
© Marc Read, Chris McEwan, Emma Hart, Ed Clark, Julie Greensmith,
Uwe Aickelin

Our group discussions concerned the derivation of a principled approach to deriving immune inspired engineering principles from immune system simulations. Drawing on the group's experience in developing complex immune simulations and attempting to create immune inspired algorithms, a principled approach to extracting from detailed simulations the minimal and essential set of principles required to capture a phenomenon of interest (PoI) was outlined.

These detailed simulations are typically developed to aid in exploring an immune domain, with no view to contribute to bio-inspired engineering solutions.

However, such simulations may be deemed to have captured some PoI with respect to engineering solutions, such as self-organisation or memory. Our approach advocates a principled approach to distilling such simulations to their *minimal representations*; the simulation may capture a great deal of immunological detail that is not essential to the PoI.

The *minimal representation* represents the smallest set of entities and their interactions required to capture the PoI. Distilling a detailed simulation to the minimal representation requires that one establish tests to demonstrate the presence of the PoI; the distillation process entails iteratively stripping complexity from the original simulation to the point that the PoI can no longer be maintained. Sensitivity analysis is identified as a means of indicating which elements of the simulation may be stripped; by attributing variation of a system's outputs to variation in its inputs, sensitivity analysis can reveal which cells and molecules of the system are non-critical with respect to the PoI.

Our approach is not intended as a replacement of the conceptual framework [1], or "immuno-engineering" [2]. Rather it elucidates some of the processes expressed within these concepts, and how the parties within and around the field of artificial immune systems relate to the provision of bio-inspired engineering solutions.

References

- 1 Susan Stepney, Robert E. Smith, Jonathan Timmis, Andy M. Tyrrell, Mark J. Neal and Andrew N. W. Hone. *Conceptual Frameworks for Artificial Immune Systems*. International Journal of Unconventional Computing, 1(3):315-338, 2005.
- 2 J. Timmis, E. Hart, A. Hone, M. Neal, A. Robins, S. Stepney and A. Tyrrell. *Immuno-Engineering*. 2nd IFIP International Conference on Biologically Inspired Collaborative Computing, 20th IFIP World Computer Congress, Milan, Italy, September 2008. IEEE Press Vol: 268/2008 pp. 3-17. 2008.

Participants

- Uwe Aickelin
University of Nottingham, GB
- Luca Albergante
Università di Milano, IT
- Bruno Apoloni
Università di Milano, IT
- Helio J.C. Barbosa
Lab. Nacional de Computação Científica-Petrópolis, BR
- Ed Clark
University of York, GB
- George M. Coghill
University of Aberdeen, GB
- Benjamin Doerr
MPI für Informatik – Saarbrücken, DE
- Julie Greensmith
University of Nottingham, GB
- Emma Hart
Edinburgh Napier University, GB
- Christian Jacob
University of Calgary, CA
- Thomas Jansen
University College Cork, IE
- Per Kristian Lehre
Technical Univ. of Denmark, DK
- Chris McEwan
Edinburgh Napier University, GB
- Yevgen Nebesov
Leibniz Univ. Hannover, DE
- Robert Oates
University of Nottingham, GB
- Pietro Oliveto
University of Birmingham, GB
- Mathias Pacher
Univ. Frankfurt am Main, DE
- Mark Read
University of York, GB
- Thomas Stibor
TU München, DE
- Dirk Sudholt
University of Birmingham, GB
- Johannes Textor
Universität Lübeck, DE
- Jon Timmis
University of York, GB
- Alan FT Winfield
University of the West of England – Bristol, GB
- Carsten Witt
Technical Univ. of Denmark, DK
- Lidia Yamamoto
Université de Strasbourg – Strasbourg, FR
- Christine Zarges
TU Dortmund, DE

