DAGSTUHL
REPORTS

Volume 3, Issue 6, June 2013

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
October, 2013

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`.

Digital Object Identifier: 10.4230/DagRep.3.6.i

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.

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

Report from Dagstuhl Seminar 13231

Belief Change and Argumentation in Multi-Agent Scenarios

Edited by

Jürgen Dix[1], Sven Ove Hansson[2], Gabriele Kern-Isberner[3], and Guillermo Simari[4]

1 **TU Clausthal, DE,** dix@tu-clausthal.de
2 **KTH Stockholm, SE,** soh@kth.se
3 **TU Dortmund, DE,** gabriele.kern-isberner@cs.uni-dortmund.de
4 **Universidad Nacional del Sur – Bahia Blanca, AR,** gisimari@gmail.com

Abstract

This report documents the programme and outcomes of Dagstuhl Seminar 13231 "Belief Change and Argumentation in Multi-Agent Scenarios". The seminar brought together researchers from the fields of argumentation theory and belief revision, both from philosophy and computer science, to present recent research results and exchange ideas for combining argumentation and belief revision. A key objective of the seminar, moreover, has been to shed light on the applicability of these two fields in the area of multi-agent systems: Can both argumentation and belief revision be combined and used in a rational agent? Before revising its beliefs, how should an agent decide *what*, or *what part of the new information* should be believed? Can this deliberation before the proper revision process be performed by argumentation?

The unique atmosphere of Dagstuhl provided again a perfect environment for leading researchers from a wide variety of backgrounds to discuss future directions of argumentation, belief revision and their applications in multi-agent systems.

Seminar 03.–07. June, 2013 – www.dagstuhl.de/13231
1998 ACM Subject Classification I.2.4 Knowledge Representation Formalisms and Methods
Keywords and phrases Belief revision, argumentation, multi-agent systems
Digital Object Identifier 10.4230/DagRep.3.6.1
Edited in cooperation with Matthias Thimm

1 Executive Summary

Jürgen Dix
Sven Ove Hansson
Gabriele Kern-Isberner
Guillermo Simari

Belief change and argumentation theory both belong to the wide field of knowledge representation, but their focal points are different. Argumentation theory provides frameworks for reasoning by setting up formal structures that allow the processing and evaluation of arguments for or against a certain option. Here, focus is put on dialectical deliberation and on finding justifications for decisions. Belief change theory has its focus on the adjustments of previously held beliefs that are needed in such processes. However, the interrelations between the two fields are still for the most part unexplored.

Both the fields of argumentation theory and belief revision are of substantial relevance for multi-agent systems which are facing heavy usage in industrial and other practical applications in diverse areas, due to their appropriateness for realizing distributed autonomous systems. Moreover, the topics of this seminar address recent research questions in the general area of decision making and are innovative in the combination of methods.

The seminar took place June 3rd–7th 2013, with 39 participants from 16 countries. The program included overview talks, individual presentations by the participants and group discussions. Overview talks ranged from 30 to 35 minutes, individual presentations were about 25 minutes long, including questions. We specifically asked participants not to present current research (their next conference paper), but rather asked to relate their research to argumentation/belief revision and how it could be used in agent theories.

Participants were encouraged to use their presentations to provide input for the discussion groups. We organized two discussion groups that each met twice (they took place in the afternoon, before and after the coffee break). Each group was headed by two organizers as discussion leaders (see Section 4).

The seminar concluded with the presentation of the group discussions on Friday morning and a wrap-up of the seminar.

From the discussion groups, some core topics arose which will help to focus further scientific work: Semantical issues concerning belief revision and argumentation were seen to be of major importance, and a layered view on both argumentation and belief revision, separating the underlying logic from the argumentation layer resp. revision layer helped to provide common grounds for the two communities. Both these topics proved to be very successful to stimulate scientific discourse, gave rise to interesting questions that might lead to papers and projects in the future, and look promising to allow a deeper analysis and a better understanding of the links between the two areas. Furthermore, a strong interest in having more applications and benchmarks became obvious, and a road map collecting informations on that is planned.

The organizers agreed to put together a special issue of *Annals of Mathematics and Artificial Intelligence* on *Argumentation and Belief revision* and invite papers on the use of methods and tools from belief change theory in argumentation theory, on the use of methods and tools from argumentation theory in belief change theory, on systems and frameworks that contain elements from both belief change and argumentation, and on practical applications of argumentation or belief revision in multi-agent systems or knowledge representation.

2 Table of Contents

3 Overview of Talks

3.1 Pareto Optimality and Strategy Proofness in Group Argument Evaluation

Edmond Awad (Masdar Institute – Abu Dhabi, AE)

Joint work of Edmond Awad, Mikolaj Podlaszewski, Martin Caminada, Gabriella Pigozzi

Collective argument evaluation is the problem of aggregating multiple opinions about how a given set of arguments should be evaluated. However, finding a consistent collective evaluation might not be the only concern. Two key criteria, to consider, are Pareto optimality and strategy proofness, which are fundamental in any social choice and multi-agent setting. Two aggregation operators were studied with respect to Pareto optimality and strategy proofness. However, these studies were built on naive models of preferences. In this study, we propose more realistic types of preferences and use them to study three operators from literature with respect to Pareto optimality and strategy proofness.

3.2 On the Maximal and Average Numbers of Stable Extensions

Ringo Baumann (Universität Leipzig, DE)

Joint work of Baumann, Ringo; Strass, Hannes
Main reference R. Baumann, H. Strass, "On the Maximal and Average Numbers of Stable Extensions," in Proc. of the Second Int'l Workshop on Theory and Applications of Formal Argumentation (TAFA'13), Aug. 2013.

We present an analytical and empirical study of the maximal and average numbers of stable extensions in abstract argumentation frameworks. As one of the analytical main results, we prove a tight upper bound on the maximal number of stable extensions that depends only on the number of arguments in the framework. More interestingly, our empirical results indicate that the distribution of stable extensions as a function of the number of attacks in the framework seems to follow a universal pattern that is independent of the number of arguments.

The obtained results can be used to provide lower bounds for the minimal realizibility of certain sets of extensions. Furthermore, counting techniques may yield upper bounds for algorithms computing extensions. Finally, the average number gives some guidance on how many extensions a given AF with n arguments and m attacks will have.

3.3 Changes driven by goals in argumentation: Framework and tool

Pierre Bisquert (Paul Sabatier University – Toulouse, FR)

Joint work of Bisquert, Pierre; Dupin de St-Cyr, Florence; Cayrol, Claudette; Lagasquie, M.C.

This work defines a new framework for dynamics in argumentation. In this framework, an agent can change an argumentation system (the target system) in order to achieve some desired goal. Changes consist in addition/removal of arguments or attacks between arguments. The agent must respect some constraints induced by her own knowledge encoded by another argumentation system. We present a software that computes the possible change operations for a given agent on a given target argumentation system in order to achieve some given goal.

3.4 Toward a Constructive Theory of Epistemic Change

Alexander Bochman (Holon Institute of Technology, IL)

We discuss some directions and desiderata for the development of a constructive theory of epistemic change, paying special attention to the representation problem of distributed knowledge, as well as the relationships of belief change to inference and argumentation.

3.5 Argumentation as Inference vs Argumentation as Dialogue

Martin Caminada (University of Aberdeen, GB)

Joint work of Caminada, Martin; Podlaszewski, Mikołaj
Main reference M. Caminada, M. Podlaszewski, "Grounded Semantics as Persuasion Dialogue," in Proc. of Computational Models of Arguments (COMMA'12), Frontiers in Artificial Intelligence and Applications, Vol. 245, pp. 478–485, IOS Press, 2012.
URL http://dx.doi.org/10.3233/978-1-61499-111-3-478
URL http://homepages.abdn.ac.uk/martin.caminada/pages/publications/COMMA_grounded_game.pdf

In the formal argumentation community, one can distinguish two main lines of research: argumentation as inference and argumentation as dialogue. The first line of research, going back to the work of Pollock, Vreeswijk and Simari & Loui, is focused on argumentation as a way of performing non-monotonic entailment. That is, it is focused on the *outcome* of argumentation. The second line of research, going back to the work of Hamblin, Mackenzie and Walton & Krabbe, is focused on argumentation as dialectics, involving different parties. That is, it is focused on the *process* of argumentation.

In our recent work, we aim to reconcile these two lines of research. In particular, we are able to express argument-based entailment as the ability to win a particular type of discussion.

3.6 Enforcement in Argumentation is a kind of Update

Florence Dupin de St-Cyr (Paul Sabatier University – Toulouse, FR)

Joint work of Bisquert, Pierre; Cayrol, Claudette; Dupin de Saint-Cyr, Florence; Lagasquie-Schiex, Marie-Christine
Main reference P. Bisquert, C. Cayrol, F. Dupin de Saint-Cyr, M.-C. Lagasquie-Schiex, "Enforcement in Argumentation is a kind of Update," in Proc. of the 7th Int'l Conf. on Scalable Uncertainty Management (SUM'13), Washington DC, USA, September 16 to 18, 2013.
URL http://www.irit.fr/publis/ADRIA/PapersDupin/Draftsum2013.pdf

In the literature, enforcement consists in changing an argumentation system in order to force it to accept a given set of arguments. In this paper, we extend this notion by allowing incomplete information about the initial argumentation system. Generalized enforcement is an operation that maps a propositional formula describing a system and a propositional formula that describes a goal, to a new formula describing the possible resulting systems. This is done under some con- straints about the allowed changes. We give a set of postulates restraining the class of enforcement operators and provide a representation theorem linking them to a family of proximity relations on argumentation systems.

3.7 Potential knowledge

Andre Fuhrmann (Goethe-Universität Frankfurt am Main, DE)

The thesis that every truth is knowable is usually glossed by decomposing knowability into possibility and knowledge. Under elementary assumptions about possibility and knowledge, considered as modal operators, the thesis collapses the distinction between truth and knowledge (as shown by the so-called Fitch-argument). As far as a purely logical refutation of the knowability thesis comes as a surprise, the Fitch-argument is paradoxical.—We show that there is a more plausible way of interpreting knowability such that the Fitch-argument does not apply. In this interpretation possibility acts not as a modal operator but as a modal modifier of a modal operator. We call this the potential knowledge-interpretation of knowability. We compare our interpretation with the rephrasal of knowability proposed by Edgington and Rabinowicz & Segerberg, inserting an actuality-operator. This proposal shares some key features with ours but suffers from requiring specific transworld-knowledge. We observe that potential knowledge involves no transworld-knowledge. We describe the logic of potential knowledge by providing a new type of models, hyperrelational models, for interpreting the new operator. Finally we show that potential knowledge cannot be fitched: The knowability thesis can be added to the elementary logic of potential knowledge without collapsing modal distinctions.

3.8 Identity Merging and Identity Revision in Talmudic Logic

Dov Gabbay (King's College London, GB)

Let K be a classical monadic theory (let us call it the *Surface* theory) and let a and b be two constants. The theory might say $P(a) \wedge \neg P(b)$. Suppose we add the identity input $a = b$. We get inconsistency. Ordinary AGM revision will do whatever it does to restore consistency. AGM is not adequate in normative legal context

However, the legal system of the Talmud does it differently. It will ask , where do $P(a)$ and $\neg P(b)$ come from (i.e. how are they derived from other basic principles; why do we have them in the theory?) and once we know that we can decide which one of the two to choose. So we need to assume the existence of another *deep* theory , from which the sentence of the *surface* theory are derived non-monotonically. With each surface monadic predicate $P(c)$ we associate a deep base theory $\Delta[P(c)]$ which derives either $P(c)$ or $\neg P(c)$. When we revise by $a = b$, we look at $\Delta[P(a)] \cup \Delta[P(b)]$. This theory might prove different results, being non-monotonically more specific. For example take the recent Boston marathon terrorist problem. The law deals one way with the rights of terrorists and allows sending them to a prison camp $P(a)$ and another way with the rights of american citizens, not allowing sending them to prison camp $\neg P(b)$. How to deal with entities which are both? (i.e deal with the input $a = b$?) We ask why do we have $P(a)$? The answer is that there is the deep theory of *security of the nation*. Why do we have $\neg P(b)$? The answer is that there is the deep theory of *human rights*. Taking the union of these two deep theories and see what we can derive.

This model is *good for legal reasoning* but is not adequate for Talmudiic reasoning. To model Talmudic legal reasoning we need to use logics with defeaters or reactive cancellations. Both $\neg P(a)$ and $\neg P(b)$ can be derived but security cancels the negation $\neg P(a)$, and we end up with $P(a) \wedge \neg P(b)$. When we have the input $a = b$, then the security cancellation the negation $\neg P(a)$ is itself cancelled and we end up with $\neg P(a) \wedge \neg P(b)$. So we need a non monotonic theory of reactive cancellations of any level which can be used to implement revision of the form $x = y$.

3.9 Mixed-initiative argumentation and Arguing with optimization norms

Aditya K. Ghose (University of Wollongong, AU)

Main reference A.K. Ghose, B.T.R. Savarimuthu, "Norms as Objectives: Revisiting Compliance Management in Multi-Agent Systems," in Proc. of the 14th Int'l Workshop on Coordination, Organisations, Institutions and Norms (COIN'12), LNCS, Vol.7756, pp. 105–122, Springer, 2012.
URL http://dx.doi.org/10.1007/978-3-642-37756-3_7
Main reference C. Fon Chang, A.K. Ghose, A. Miller, "Mixed-initiative argumentation: A framework for justification management in clinical group decision support," in Proc. of the AAAI 2009 Fall Symp. on the Uses of Computational Argument, AAAI, 2009.
URL http://aaai.org/ocs/index.php/FSS/FSS09/paper/view/982
Main reference C. Fon Chang, A. Miller, A.K. Ghose, "Mixed-Initiative Argumentation: Group Decision Support in Medicine," in Proc. of the Second Int'l ICST Conf. on Eletronic Healthcare (eHealth'09), Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, Vol. 27, pp. 43–50, Springer, 2010.
URL http://dx.doi.org/10.1007/978-3-642-11745-9_8

In the first part of this talk, I will explore a novel integration of belief revision and argumentation in the following setting. Imagine a series of rounds of argumentation where, say, a

group of experts bring to bear the same body of knowledge (as represented by the arguments they articulate and the preferences they apply) to a sequence of decisions. It is important to ensure that this body of knowledge (and in particular the set of preferences) is consistently applied in this sequence of decisions. This is also important in decision *forensics*, where this decision sequence is audited to identify errors or inconsistencies. Imagine the ability to *invert* the conventional argumentation machinery so that we are able to present a set of *winning* arguments and ask how the background theory (the set of arguments and preferences that have been brought to bear in prior decisions) might be minimally modified so that the resulting theory would generate this precise set of arguments as the winning arguments. This machinery has been implemented in a system called JUST-CLINICAL and used in clinical quality assurance in settings where specialist oncologists argue over the best course of treatment for a patient. I will also use this talk to draw attention to other applications of such an approach, including, most importantly, the design of normative frameworks for multi-agent institutions (both electronic and otherwise).

In the second part of the talk, I will offer some preliminary results on the problem of arguing with optimization norms.

3.10 An input/output perspective on abstract argumentation in a dynamic environment

Massimiliano Giacomin (University of Brescia, IT)

Joint work of Baroni, Pietro; Boella, Guido; Cerutti, Federico; Giacomin, Massimiliano; Van Der Torre, Leendert; Villata, Serena
Main reference P. Baroni, G. Boella, F. Cerutti, M. Giacomin, L. Van Der Torre, S. Villata, "On input/output argumentation frameworks ", in Proc. of Computational Models and Arguments (COMMA'12), Frontiers in Artificial Intelligence and Applications, Vol. 245, pp. 358–365, IOS Press, 2012.
URL http://dx.doi.org/10.3233/978-1-61499-111-3-358

This talk considers the interaction between argumentation-based intelligent autonomous agents in a dynamic environment, by focusing on their characterization in terms of input/output systems. More specifically, each agent can be modelled as a Dung's partial argumentation framework exposing a well- defined external interface. Two relevant questions are: i) whether the justification status of arguments according to a given semantics can be determined on the basis of local computations on partial argumentation frameworks, and ii) whether subframeworks with the same input/output behavior can be interchanged without affecting the result of semantics evaluation of the other arguments interacting with them. In order to answer the first question, the notion of semantics decomposability is introduced and the main admissibility-based semantics are analyzed in this respect. In order to answer the second question, the talk introduces Argumentation Multipoles to characterize the behavior of an argumentation framework as a black box, and studies the interchangeability of equivalent Argumentation Multipoles under admissibility-based semantics. Some applications of these results, including the study of dynamic frameworks, are finally outlined.

3.11 Belief change for finite minds

Sven Ove Hansson (KTH Stockholm, SE)

Standard models of belief change such as partial meet contraction operate by making choices among cognitively inaccessible objects such as possible worlds or maximal consistent subsets that lack a finite representation. Finite belief bases avoid that difficulty, but bring in others. An alternative approach is presented in which changes take place on finitely representable belief sets but no distinction is made between different belief bases for the same belief set. Reference to infinite objects is avoided by changing the level of selection. Choice functions can operate directly on the set of possible outcomes (the credible and reachable finite-based belief sets) rather than on infinite and cognitively inaccessible objects.

3.12 Logics for belief change operations: a short history of nearly everything

Andreas Herzig (Paul Sabatier University – Toulouse, FR)

Main reference Journées IAF, Aix-en-Provence, June 2013

We examine several belief change operations in the light of dynamic logic of propositional assignments DL-PA. We show that we can encode in a systematic way update and revision operators (such as Winslett's PMA operator and Dalal's revision operator) as particular DL-PA programs. This provides a syntactical update method.

3.13 A Probabilistic Approac to Modelling Uncertain Logical Arguments

Anthony Hunter (University College London, GB)

Main reference A. Hunter, "A probabilistic approach to modelling uncertain logical arguments," Int'l Journal of Approximate Reasoning, 54(1):47–81, 2013.
URL http://dx.doi.org/10.1016/j.ijar.2012.08.003

Argumentation can be modelled at an abstract level using a directed graph where each node denotes an argument and each arc denotes an attack by one argument on another. Since arguments are often uncertain, it can be useful to quantify the uncertainty associated with each argument. Recently, there have been proposals to extend abstract argumentation to take this uncertainty into account. This assigns a probability value for each argument that represents the degree to which the argument is believed to hold, and this is then used to generate a probability distribution over the full subgraphs of the argument graph, which in turn can be used to determine the probability that a set of arguments is admissible or an extension. In order to more fully understand uncertainty in argumentation, in this paper, we extend this idea by considering logic-based argumentation with uncertain arguments. This is based on a probability distribution over models of the language, which can then be used to give

a probability distribution over arguments that are constructed using classical logic. We show how this formalization of uncertainty of logical arguments relates to uncertainty of abstract arguments, and we consider a number of interesting classes of probability assignments.

3.14 On the Revision of Argumentation Systems: Minimal Change of Arguments Status

Sebastien Konieczny (Artois University, FR)

Joint work of Coste-Marquis, Sylvie; Konieczny, Sebastien; Mailly, Jean-Guy; Marquis, Pierre

In this work, we investigate the revision issue for argumen- tation systems a‘ la Dung. We focus on revision as minimal change of the arguments status. Contrarily to most of the previous works on the topic, the addition of new arguments is not allowed in the revision process, so that the revised system has to be obtained by modifying the attack relation, only. We introduce a language of revision formulae which is ex- pressive enough for enabling the representation of complex conditions on the acceptability of arguments in the revised system. We show how AGM belief revision postulates can be translated to the case of argumentation systems. We provide a corresponding representation theorem in terms of minimal change of the arguments status. Several distance-based revision operators satisfying the postulates are also pointed out.

3.15 Toward Incremental Computation of Argumentation Semantics: A Decomposition-based Approach

Beishui Liao (Zhejiang University, CN)

Main reference B. Liao, "Toward Incremental Computation of Argumentation Semantics: A Decomposition-based Approach," Annals of Mathematics and Artificial Intelligence, Vol. 67, Issue 3–4, pp. 319–358, 2013.
URL http://dx.doi.org/10.1007/s10472-013-9364-8

Currently, except some classes of argumentation frameworks (with special topologies or fixed parameters, such as acyclic, symmetric, and bounded tree- width, etc.) that have been clearly identified as tractable, for a generic argumentation framework (also called a defeat graph), how to efficiently compute its semantics is still a challenging problem. Inspired by the local tractability of an argumentation framework, we first propose a decomposition- based approach, and then conduct an empirical investigation. Given a generic argumentation framework, it is firstly decomposed into a set of sub-frameworks that are located in a number of layers. Then, the semantics of an argumentation framework are computed incrementally, from the lowest layer in which each sub-framework is not restricted by other sub-frameworks, to the highest layer in which each sub-framework is most restricted by the sub- frameworks located in the lower layers. In each iteration, the semantics of each sub-framework is computed locally, while the combination of semantics of a set of sub-frameworks is performed in two dimensions: horizontally and vertically. The average results show that when the ratio of the number of edges to the number of nodes of a defeat graph is less than 1.5:1, the decomposition-based approach is obviously efficient.

3.16 Argumentation and Belief Revision in Datalog+/- Ontologies Integration

Maria Vanina Martinez (University of Oxford, GB)

Joint work of Deagustini, Cristhian A. D.; Falappa, Marcelo; Simari, Guillermo R.

The increasing number of available knowledge bases in the form of ontologies accessible online makes their integration a concrete necessity in order to fully exploit the knowledge stored in these resources. While it is possible through different means to maintain a consistent ontology, it is certainly more difficult for the answers obtained separately from consistent ontologies to remain consistent when considered together. We are working towards an approach to merge multiple Datalog+/- ontologies, a family of rule-based ontological languages, addressing the incoherence and inconsistency problems related to this process.

The main idea is to develop a kernel contraction-like approach to solve conflicts in the merging process. This approach restores coherence/consistency by applying incision functions that select formulas to delete from the minimal incoherent/inconsistent sets encountered in the union of the ontologies. We wish to consider each ontology as an independent entity that represents information locally, and use that information in order to define an adequate incision function. However, in the presence of conflicts either each ontology or a set of them together can provide arguments in favor or against the different pieces of information that add up to the conflicts, and the decision of what information to remove from consideration can be identified through an argumentation process. It is therefore a principal part of this work to extend ontological languages with argumentation capabilities in order to allow ontologies (or sets of ontologies) to supply information that can be challenged and argued for and/or against in the merging process.

3.17 Reconfiguration of Large-Scale Surveillance Systems

Peter Novak (TU Delft, NL)

Joint work of Novak, Peter; Cees Witteveen

Metis project aims at study of techniques supporting development of large-scale dependable surveillance system-of-systems for maritime safety and security. Surveillance systems, such as Metis, typically comprise a number of heterogeneous information sources and information aggregators. Among the main problems of their deployment lies scalability of such systems with respect to a potentially large number of monitored entities. One of the solutions to the problem is continuous and timely adaptation and reconfiguration of the system according to the changing environment it operates in. At any given timepoint, the system should use only a minimal set of information sources and aggregators needed to facilitate cost-effective early detection of indicators of interest.

On the background of Metis prototype description, I will introduce a theoretical framework for modelling scalable information-aggregation systems. We model such systems as networks of inter-dependent reasoning agents, each representing a mechanism for justification/refutation of a conclusion derived by the agent. The algorithm facilitating continuous reconfiguration is based on standard results from abstract argumentation and corresponds to computation of a grounded extension of the argumentation framework associated with the system.

3.18 Relaxing Independence Assumption in Probabilistic Argumentation

Nir Oren (University of Aberdeen, GB)

Joint work of Li, Hengfei; Oren, Nir; Norman, Timothy J.
Main reference H. Li, N. Oren, T.J. Norman, "Relaxing Independence Assumption in Probabilistic Argumentation," in Proc. of the 10th Int'l Workshop on Argumentation in Multiagent Systems (ArgMAS'13), Saint Paul, Minnesota, USA, May 6–10 2013, 2013.
URL http://homepages.abdn.ac.uk/n.oren/pages/publications/li13relaxing.pdf

Probabilistic argumentation frameworks (PrAFs) are a novel extension to standard argumentation systems, enabling one to reason about the likelihood of a set of arguments appearing within an extension. However, PrAFs assume that the likelihood of arguments appearing is independent of the presence of other arguments. In this paper, we lift this restriction through the introduction of probabilistic evidential argumentation frameworks (PrEAFs). Our extension captures probabilistic dependencies through the use of a support relation, as used in bipolar argumentation frameworks. After describing PrEAFs and their properties, we present algorithms for computing PrEAF semantics.

3.19 System of Spheres-based Constructions of Transitively Relational Partial Meet Multiple Contractions

Mauricio Reis (University of Madeira – Funchal, PT)

Joint work of Peppas, Pavlos; Reis, Maurício D. L.; Fermé, Eduardo
Main reference E. Fermé,M.D.L. Reis, "System of spheres-based multiple contractions," Journal of Philosophical Logic, 41(1):29–52, 2012.
URL 10.1007/s10992-011-9197-z
Main reference M.D.L. Reis, "On Theory Multiple Contraction," PhD thesis, Universidade da Madeira, May 2011.
URL http://hdl.handle.net/10400.13/255

We show that not all the system of spheres-based multiple contractions (abbrev. SS-bMCs) are transitively relational partial meet multiple contractions (abbrev. TRPMMCs) and, vice versa, that not all TRPMMCs are SS-bMCs. Furthermore, we show that, contrary to what is the case in what concerns contractions by a single sentence, there is not a system of spheres-based construction of multiple contractions which originates each and every TRPMMC. Finally, we propose two ways of generalizing Grove's system of spheres-based contractions to the case of multiple contractions, which originate only TRPMMCs.

3.20 A Logical Theory about Dynamics in Abstract Argumentation

Tjitze Rienstra (University of Luxembourg, LU)

Joint work of Booth, Richard; Kaci, Souhila; Rienstra, Tjitze; van der Torre, Leendert

We address dynamics in abstract argumentation using a logical theory where an agent's belief state consists of an argumentation framework (AF, for short) and a constraint that encodes

the outcome the agent believes the AF *should* have. Dynamics enters in two ways: (1) the constraint is strengthened upon learning that the AF should have a certain outcome and (2) the AF is expanded upon learning about new arguments/attacks. A problem faced in this setting is that a constraint may be inconsistent with the AF's outcome. We discuss two ways to address this problem: First, it is still possible to form consistent *fallback beliefs*, i.e., beliefs that are most plausible given the agent's AF and constraint. Second, we show that it is always possible to find AF expansions to restore consistency. Our work combines various individual approaches in the literature on argumentation dynamics in a general setting.

3.21 Decision Making in Knowledge Integration with Dynamic Creation of Argumentation

Ken Satoh (NII – Tokyo, JP)

Joint work of Satoh, Ken; Takahashi, Kazuko
Main reference K. Satoh, K. Takahashi, "Decision Making in Knowledge Integration with Dynamic Creation of Argumentation," in Proc. of the Int'l Workshop on Information Search, Integration and Personalization (ISIP'12) / Revised Selected Papers, CCIS, Vol. 146, pp. 41–50, Springer, 2012.
URL http://dx.doi.org/10.1007/978-3-642-40140-4_5

We discuss a semantics of dynamic creation of arguments when knowledge from different agents are combined. This arises when an agent does not know the other agent's knowledge and therefore, the agent cannot predict which arguments are attacked and which counterarguments are used in order to attack the arguments. In this paper, we provide a more general framework for such argumentation system than previous proposed framework and provide a computational method how to decide acceptability of argument by logic programming if both agents are eager to give all the arguments.

3.22 Argumentation semantics and update semantics

Jan Sefranek (University of Bratislava, SK)

Main reference unpublished

A dominance of new information is an accepted feature of updates according to Katsuno-Mendelzon postulates. It is argued in this paper that new updating information could be ignored, if it depends on (defeasible) assumptions falsified by a description of a current state. This is a feature of our approach to updates of assumption-based frameworks, presented in this paper. The approach is applicable to updates of non-monotonic knowledge bases, which can be described in terms of an assumption-based framework. The updated knowledge base may be characterized semantically using different argumentation semantics. The presented framework may be applied to some types of multi-agent scenarios.Main contributions of the paper are as follows. According to our best knowledge,this is the first paper devoted to updates of assumption-based frameworks. We distinguish between preferential conflicts solving and updating. Update-solvable conflicts are defined and only those are resolved. Updating is based on sound sets of assumptions, which are not defeated by their subsets. Argumentation semantics are applied to sound sets of assumptions and admissible,maximal

admissible (preferred), stable, complete and well-founded update operations are defined. Irrelevant updates are defined and it is shown that update operations return empty set of sets of assumptions for irrelevant updates. A proposition about the inertia of a current state was proven for stable sets of assumptions. The proposition does not hold for maximal admissible(preferred) sets of assumptions, if a non-normal assumptions-based framework is given.

3.23 Probabilistic Presumption-based Argumentation with Applications to Cyber-security

Gerardo I. Simari (University of Oxford, GB)

In cyber-security applications, intelligence information and data about past attacks is often used when trying to determine the perpetrator of a specific attack of interest—this is called the "attribution problem". Knowledge bases consisting of all the information at hand are bound to contain contradictory data coming from different sources, as well as data with varying degrees of uncertainty attached. Likewise, an important aspect of this effort is deciding what information is no longer useful: intelligence reports may be outdated, may come from sources that have recently been discovered to be of low quality, or abundant evidence may be available that contradicts them. A framework capable of providing decision support in this domain must therefore be capable of: (i) handling contradictory information; (ii) answering abductive queries; (iii) managing uncertainty; and (iv) updating beliefs. Presumptions come into play as key components of answers to abductive queries, and must be maintained as elements of the knowledge base; therefore, whenever candidate answers to these queries are evaluated, the (in)consistency of the knowledge base together with the presumptions being made needs to be addressed via belief revision operations. In this talk, we describe preliminary work in the development of a probabilistic presumption-based argumentation framework to model the processes of solving attribution problems and maintaining the knowledge bases used to do so.

3.24 A (Very) Brief (and Incomplete) Overview of Argumentation Systems

Guillermo R. Simari (National University of the South – Bahia Blanca, AR)

Main reference

he research on the theoretical foundations and practical applications of Argumentation in Artificial Intelligence has been expanding at an increasingly fast pace for the last three decades. This activity has helped in the demarcation of many subareas as the investigations gained momentum, creating a field that is exciting, fruitful, and rewarding. The challenges are many, and they are met with methods and techniques that have enriched the area of Knowledge Representation and Reasoning. In this tutorial, a short structured overview of the area of Argumentation Systems will be provided in order to lay a foundation for further

discussion. We will present the intuitions and fundamentals of the process of argumentation together with a succinct introduction to abstract argumentation systems and the elements of four different of systems of concrete argumentation where arguments are constructed from a knowledge base.

References

1 Besnard, P., Hunter, A.: A Logic-Based Theory of Deductive Arguments. Artif. Intell. 128(1-2), 203–235 (2001).
2 Bondarenko, A., Dung, P. M., Kowalski, R., Toni, F.: An abstract, argumentation-theoretic approach to default reasoning. Art. Int., 93(1- 2):63–101, 1997.
3 Dung, P.M.: On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. Artificial Intelligence 77(2), 321–358 (1995).
4 García, A.J., Simari, G.R.: Defeasible logic programming: An argumentative approach. Theory and Practice of Logic Programming 4(1-2), 95–138 (2004).
5 Prakken, H.: An abstract framework for argumentation with structured arguments. Argument and Computation 1, 93–124 (2009).

3.25 On Stratified Labelings for Argumentation Frameworks and Ranking Functions

Matthias Thimm (Universität Koblenz-Landau, DE)

Joint work of Thimm; Matthias, Kern-Isberner, Gabriele
Main reference M. Thimm, G. Kern-Isberner, "Stratified Labelings for Abstract Argumentation," Preliminary Report, arXiv:1308.0807v1 [cs.AI], 2013.
URL http://arxiv.org/abs/1308.0807v1

We introduce stratified labelings as a novel semantical approach to abstract argumentation frameworks. Compared to standard labelings, stratified labelings provide a more fine-grained assessment of the status of arguments using ranks instead of the usual labels "in", "out", and "undecided". We relate the framework of stratified labelings to conditional logic and, in particular, to the System Z ranking functions.

3.26 Possibilistic Belief Revision with Fuzzy Argumentation based on Trust

Serena Villata (INRIA Sophia Antipolis – Méditerranée, FR)

Joint work of da Costa Pereira, Celia; Tettamanzi, Andrea; Villata, Serena
Main reference C. da Costa Pereira, A. Tettamanzi, S. Villata, "Changing One's Mind: Erase or Rewind? Possibilistic Belief Revision with Fuzzy Argumentation Based on Trust," in Proc. of the 22nd Int'l Joint Conf. on Artificial Intelligence (IJCAI'11), pp. 164–171, IJCAI/AAAI, 2011.
URL http://ijcai.org/papers11/Papers/IJCAI11-039.pdf

Belief revision aims at describing the changes in the agents mind in response to new information. On the other hand, one of the important concerns in argumentation is the strategies employed by an agent in order to succeed in persuading other agents to change their

mind. In this talk, we will present our ongoing work on combining these two complementary fields of research into a unitary multi-agent framework, and we will highlight the future research lines we are pursuing. In this framework, each piece of information is represented as an argument which can be more or less accepted depending on the trustworthiness of the agent who proposes it. We adopt possibility theory to represent uncertainty about incoming information, and to model the fact that information sources can be only partially trusted. The three main ingredients of this framework are:

1. A fuzzy extension of the notion of argumentation framework, where arguments have a "strength" that depends on the trustworthiness degree of their sources;
2. A fuzzy labeling algorithm, which computes the degree of acceptability of each argument;
3. A mechanism to translate the computed fuzzy labeling into a possibilistic set of beliefs, whereby an agent will believe the conclusions of the accepted arguments, as well as their consequences.

The following are some of the advantages of such a framework: (i) partially trusted input can be taken into account naturally; (ii) arguments are never lost when contradictory information arrives; (iii) argument reinstatement is automatically mirrored in belief reinstatement; (iv) this also solves the "drowning" problem of (possibilistic) iterated belief revision.

3.27 Semantic instantiations of abstract argumentation

Emil Weydert (University of Luxembourg, LU)

Main reference Emil Weydert, "On the Plausibility of Abstract Arguments," in Proc. of the 12th Europ. Conf. on Symbolic and Quantitative Approaches to Reasoning with Uncertainty (ECSQARU'13), LNCS, Vol. 7958, pp. 522–533, Springer, 2013.
URL http://dx.doi.org/10.1007/978-3-642-39091-3_44

In recent years, the question of how to instantiate or interpret abstract argumentation frameworks, or how to justify different ways to evaluate arguments in the context of a given framework (like extension functions), has received increasing attention. Given the defeasible character of real life argumentation, arguments are instantiated in particular by defeasible inference trees or premise-conclusion pairs. However, most of these proposals have been inspired by consistency-based accounts of default reasoning which are known to violate desirable principles and benchmark examples.

The first goal of our work has therefore been to provide an alternative interpretation of abstract argumentation frameworks based on the ranking measure paradigm. Ranking measures are well-behaved (im)plausibility valuations offering a reasonable independence concept and a powerful semantics for default conditionals. They go back to Spohn's ranking functions, which he introduced to model the dynamics of graded plain belief. The idea is to associate with each framework a generic conditional knowledge base (i.e. ranking measure constraints) obtained by translating individual arguments and attack links into specific conditionals reflecting their intended meaning while minimizing informational commitments.

In a second step, we use ranking choice functions known from ranking-based default inference to pick up a natural canonical ranking model of a framework-induced default base, notably the JZ-ranking-model. System JZ appears to represent the best proxy of entropy maximization at the ranking level and satisfies many desiderata for defeasible inference. A set E of abstract arguments can be described by the conjunction of the acceptance and violation domains of those conditionals which correspond to arguments in, resp. not, in E.

The JZ-extensions are those E whose characterizing propositions are the most plausible, i.e. are ranked lowest by the JZ-ranking. The resulting extension semantics behaves as intended for many standard examples, but it fails to validate full reinstatement and diverges from all the known competitor semantics. It may be interesting to see whether recent ranking-based approaches to conditional revision can thereby also be exploited for argumentation dynamics.

To summarize, the ranking-semantic instantiation of abstract argumentation looks promising but requires further investigation.

3.28 On the Limits of Expressiveness in Abstract Argumentation Semantics

Stefan Woltran (TU Wien, AT)

Joint work of Dunne, Paul; Dvorak, Wolfgang; Linsbichler, Thomas

The study of extension-based semantics within the seminal abstract argumentation model of Dung has largely focused on definitional, algorithmic and complexity issues. In contrast, matters relating to comparisons of representational limits, in particular, the extent to which given collections of extensions are expressible within the formalism, have been underdeveloped. As such, little is known concerning conditions under which a candidate set of subsets of arguments are "realistic" in the sense that they correspond to the extensions of some argumentation framework for a semantics of interest. In this work, we present a formal basis for examining extension-based semantics in terms of the sets of extensions that these may express within a single framework and provide a number of characterization theorems which guarantee the existence of argumentation frameworks whose set of extensions satisfy specific conditions. We also discuss how our result apply to problems of belief change in argumentation and how they can be exploited in systems implementing abstract argumentation.

3.29 Argument Rejection and Acceptance Through Attack Abstractions

Zhiqiang Zhuang (Griffith University – Brisbane, AU)

We consider the dynamics of Dung's argumentation frameworks under preferred semantics. In particular, we define two operations, namely argument rejection and argument acceptance. Argument rejection changes an argument's status from accepted to rejected and argument acceptance changes an argument's status from rejected to accepted. The changes are achieved by removing a minimal number of attacks from the argumentation framework. Argument games are used in identifying such attacks.

4 Working Groups

4.1 Results of Discussion Group I – Semantical issues and models in BR and Argumentation

The following research topics have been discussed:

- Belief Revision and Argumentation as alternative approaches to model (human) reasoning – common (semantical) grounds and differences
- Using argumentation to do belief revision and belief revision as an input to the argumentation process
- Relationships between similar structures in belief revision and argumentation
- Different layers both in BR and Argumentation:
  - Relationship between underlying logic and argumentation layer resp. revision layer
  - Better understanding of the links between the two layers, and what happens under change processes
  - Bringing the meta level into the object level (and vice versa)
  - E.g., preferences need deeper knowledge, requiring argumentation?
- Find new semantics for belief revision and argumentation
- Effects of valuations on belief revision and argumentation (multi-valued logic)
- Probability in BR and Argumentation
- Ontological aspects of argumentation and belief revision
- Decomposability, local and global models
- Paraconsistent reasoning, voting, inconsistency resolution and application to belief revision
- Belief revision and argumentation in dialogue (Negotiation and belief revision); motivation / aim for approach; (dynamics and change)
- Relation between argumentation and non-monotonic reasoning
- Influence of context on belief revision and argumentation
- Argumentation and belief revision for observing / modelling real world decision making (processes) and for general applications

All's well that ends well:

- At the beginning of the discussion: "Belief revision and argumentation are as orthogonal as food and love – you need both, but they are different!"
- Towards the end of the discussion group: "Good food leads to good love!"

4.2 Results of Discussion Group II – Belief Revision and Argumentation: who can benefit how?

It is clear that Belief Revision (BR) and Argumentation (ARG) are complementarity. BR's goal is to maintain a consistent knowledge base (KB), while ARG "tolerates" an inconsistent KB by obtaining useful conclusions without changing it. The methods developed in BR for performing the work necessary to effect a revision over a KB involve an intermediate state where the KR possibly becomes inconsistent and at that point ARG could become useful. From the point of view of ARG, the tools developed to handle an inconsistent KB usually appeal to some form of dialectical analysis. The set of arguments that support the outcome of all the pieces of knowledge that are able to "survive" the dialectical analysis represents a consistent KB. Sometimes is necessary to deactivate arguments to change this outcome and at that point could be necessary to resort to the use of some BR techniques.

Regarding applications of these two formalism, the group stressed the importance of going beyond the abstract to concrete examples. Several areas where discussed; among others online debate and semantic web applications, diagnostic systems, handling trust, managing persuasion dialogues were mentioned.

There was also an interesting discussion over the possibility of maintaining a set of benchmarks to stimulate research, make meaningful the comparison between different implementations, and learn from the experimentation. Some members of the group felt that both areas lack a sufficient number of implemented systems, while others supported the idea that benchmarks could stimulate the development of those systems.

Another interesting issue in comparing the two fields was that meanwhile researchers in ARG are keenly interested in finding applications, that drive seems to be absent in the BR camp. Again this led to a discussion over the goals of combining the results of research in concrete applications.

Participants

- Edmond Awad
Masdar Inst. – Abu Dhabi, AE
- Pietro Baroni
University of Brescia, IT
- Ringo Baumann
Universität Leipzig, DE
- Pierre Bisquert
Paul Sabatier University – Toulouse, FR
- Alexander Bochman
Holon Institute of Technology, IL
- Martin Caminada
University of Aberdeen, GB
- Célia da Costa Pereira
University of Nice, FR
- Jürgen Dix
TU Clausthal, DE
- Florence Dupin de St-Cyr
Paul Sabatier University – Toulouse, FR
- André Fuhrmann
Goethe-Universität Frankfurt am Main, DE
- Dov M. Gabbay
King's College London, GB
- Aditya K. Ghose
University of Wollongong, AU
- Massimiliano Giacomin
University of Brescia, IT
- Sven Ove Hansson
KTH Stockholm, SE
- Andreas Herzig
Paul Sabatier University – Toulouse, FR
- Anthony Hunter
University College London, GB
- Gabriele Kern-Isberner
TU Dortmund, DE
- Sebastien Konieczny
Artois University, FR
- Patrick Krümpelmann
TU Dortmund, DE
- Daniel Lehmann
The Hebrew University of Jerusalem, IL
- Beishui Liao
Zhejiang University, CN
- Pierre Marquis
Artois University, FR
- Maria Vanina Martinez
University of Oxford, GB
- Peter Novak
TU Delft, NL
- Nir Oren
University of Aberdeen, GB
- Odile Papini
Aix-Marseille University, FR
- Matei Popovici
TU Clausthal, DE
- Mauricio Reis
Univ. of Madeira – Funchal, PT
- Tjitze Rienstra
University of Luxembourg, LU
- Ken Satoh
NII – Tokyo, JP
- Jan Sefranek
University of Bratislava, SK
- Gerardo I. Simari
University of Oxford, GB
- Guillermo R. Simari
National University of the South – Bahia Blanca, AR
- Andrea Tettamanzi
University of Nice, FR
- Matthias Thimm
Universität Koblenz-Landau, DE
- Serena Villata
INRIA Sophia Antipolis – Méditerranée, FR
- Emil Weydert
University of Luxembourg, LU
- Stefan Woltran
TU Wien, AT
- Zhiqiang Zhuang
Griffith Univ. – Brisbane, AU

Report from Dagstuhl Seminar 13232

Indexes and Computation over Compressed Structured Data

Edited by

Sebastian Maneth[1] and Gonzalo Navarro[2]

1 **Universität Leipzig, DE,** sebastian.maneth@gmail.com
2 **University of Chile – Santiago, CL,** gnavarro@dcc.uchile.cl

Abstract

This report documents the program and the outcomes of Dagstuhl Seminar 13232 "Indexes and Computation over Compressed Structured Data".

Seminar 3.–7. June, 2013 – www.dagstuhl.de/13232
1998 ACM Subject Classification E.1 Data Structures, E.2 Data storage representations, E.4 Coding and Information Theory: data compaction and compression
Keywords and phrases Compression, Indexes, Data Structures
Digital Object Identifier 10.4230/DagRep.3.6.22
Edited in cooperation with Patrick Nicholson

1 Executive Summary

Sebastian Maneth
Gonzalo Navarro

The Dagstuhl Seminar "Indexes and Computation over Compressed Structured Data" took place from June 2nd to 7th, 2013. The aim was to bring together researchers from various research directions of compression and indexing of structured data. Compression, and the ability to compute directly over compressed structures, is a topic that is gaining importance as digitally stored data volumes are increasing at unprecedented speeds. Of particular interest is the combination of compression schemes with indexes that give fast access to particular operations. The seminar was meant to inspire the exchange of theoretical results and practical requirements related compression and indexing. These points were addressed in particular

- Tractability versus Intractability for Algorithmic Problems on Compressed Data
- Compression Algorithms for Strings, Trees, and Graphs
- Indexes for Compressed Data
- Algorithms for Compressed Data
- Better Search Results: Ranking and TF/IDF
- Applications of Structure Compression to other Areas

The seminar fully satisfied our expectations. The 34 participants from 11 countries (Canada, Chile, Denmark, Finland, Germany, Great Britain, Italy, Israel, Japan, Spain, and US) had been invited by the organizers to give survey talks about their recent research related to the topic of the seminar. The talks covered topics related to compression (e.g.

grammar-based string compression) databases (e.g., XML, and top-k query answering), data structures (e.g. wavelet tries), string matching, and ranged to broad application areas such as biology. Most talks were followed by lively discussions. Smaller groups formed naturally which continued these discussions later.

We thank Schloss Dagstuhl for the professional and inspiring atmosphere. Such an intense research seminar is possible because Dagstuhl so perfectly meets all researchers' needs. For instance, elaborate research discussions in the evening were followed by local wine tasting or by heated sauna sessions.

2 Table of Contents

3 Overview of Talks

3.1 Tree Compression with Top Trees

Philip Bille (Technical University of Denmark – Lyngby, DK)

Joint work of Bille, Philip; Gørtz, Inge Li; Landau, Gad M.; Weimann, Oren
Main reference P. Bille, I.L. Gørtz, G.M. Landau, O. Weimann, "Tree Compression with Top Trees," in Proc. of the 40th Int'l Colloquium on Automata, Languages, and Programming (ICALP'13), LNCS, Vol. 7965, pp. 160–171, Springer, 2013.
URL http://dx.doi.org/10.1007/978-3-642-39206-1_14
URL http://arxiv.org/abs/1304.5702

We introduce a new compression scheme for labeled trees based on top trees. Our compression scheme is the first to simultaneously take advantage of internal repeats in the tree (as opposed to the classical DAG compression that only exploits rooted subtree repeats) while also supporting fast navigational queries directly on the compressed representation. We show that the new compression scheme achieves close to optimal worst-case compression, can compress exponentially better than DAG compression, is never much worse than DAG compression, and supports navigational queries in logarithmic time.

3.2 Updates in compressed text and compressed XML

Stefan Boettcher (Universität Paderborn, DE)

Joint work of Boettcher, Stefan; Bueltmann, Alexander; Hartel, Rita; Schluessler, Jonathan
URL http://www.cs.uni-paderborn.de/fileadmin/Informatik/AG-Boettcher/Papers_for_Download/dagstuhl2013-3.pdf

Compressed XML files and compressed column-oriented main memory text databases require to support insert and delete operations of the Nth word of the represented text T on a compressed version C(T) of T without full decompression of C(T). We present IRT (=Indexed Reversable Transformation), a block-sorting technique that differs from BWT by sorting word delimiters according to their occurence in the given text T, and we present a compression technique for IRT transformed text, such that both techniques together transform a given text T into a compressed form C(T) that allows to delete the Nth word of T from C(T) and to insert a word as the Nth word of T into C(T) without full decompression of C(T). Thereby, we enable faster delete and insert operations on compressed XML files and on compressed main memory text databases. This talk is based on a conference paper [1].

References

1 Stefan Böttcher, Alexander Bültmann, Rita Hartel, Jonathan Schlüßler: *Implementing efficient up-dates in compressed big text databases*. In Proc. 24th International Conference on Data-base and Expert Systems Applications (DEXA 2013), Prague, Czech Republic, (2013).

3.3 Grammar Indexes and Document Listing

Francisco Claude (University of Waterloo, CA)

Joint work of Munro, J. Ian; Navarro, Gonzalo;

We introduce the first grammar-compressed representation of a sequence that supports searches in time that depends only logarithmically on the size of the grammar. Given a text $T[1..u]$ that is represented by a (context-free) grammar of n (terminal and nonterminal) symbols and size N (measured as the sum of the lengths of the right hands of the rules), a basic grammar-based representation of T takes $N \lg n$ bits of space. Our representation requires $2N \lg n + N \lg u + \epsilon(n \lg n) + o(N \lg n)$ bits of space, for any $0 < \epsilon \leq 1$. It can find the positions of the occ occurrences of a pattern of length m in T in $O((m^2/\epsilon) \lg((\lg u)/(\lg n)) + (m + \text{occ}) \lg n)$ time, and extract any substring of length l of T in time $O(l + h \lg(N/h))$, where h is the height of the grammar tree.

We also show a practical version of this index adapted to solve the document listing problem on versioned documents. Our index is the first one based on grammar-compression. This allows for good results on repetitive collections, whereas classical techniques cannot achieve competitive space for solving the same problem. As a result of this, our index is about 16 times smaller when compared to the state of the art for document listing [Navarro, Puglisi, and Valenzuela, SEA2011]. Our query times are competitive with the state of the art.

3.4 LZ-Compressed String Dictionaries

Johannes Fischer (KIT – Karlsruhe Institute of Technology, DE)

Joint work of Arz, Julian; Fischer, Johannes
Main reference J. Arz, J. Fischer, "LZ-Compressed String Dictionaries," arXiv:1305.0674v1 [cs.DS], 2013.
URL http://arxiv.org/abs/1305.0674v1

We review existing compressed string dictionaries (see bibliography) and further show how to compress string dictionaries using the Lempel-Ziv (LZ78) data compression algorithm. Our approach is validated experimentally on dictionaries of up to 1.5 GB of uncompressed text. We achieve compression ratios often outperforming the existing alternatives, especially on dictionaries containing many repeated substrings. Our query times remain competitive.

Presenting the paper at Dagstuhl proved to be very fruitful, because several participants noted that if our LZ-parsing is done in reverse direction and the resulting phrases are subsequently reversed again, then it is possible to relate the size of the resulting data structures to the number of phrases in the original LZ78 parsing. Thanks a lot!

3.5 A Faster Grammar-Based Self-Index

Travis Gagie (University of Helsinki, FI)

Joint work of Gagie, Travis; Gawrychowski, Paweł; Karkkainen, Juha; Nekrich, Yakov; Puglisi, Simon
Main reference T. Gagie, P. Gawrychowski, J. Kärkkäinen, Y. Nekrich, S.J. Puglisi, "A Faster Grammar-Based Self-Index," arXiv:1109.3954v6 [cs.DS], 2013; randomized version submitted to "Information and Computation".
URL http://arxiv.org/abs/1109.3954v6

To store and search genomic databases efficiently, researchers have recently started building compressed self-indexes based on grammars and LZ77. We show how, given a text $S[1..n]$ whose LZ77 parse consists of z phrases, we can build an $O(z \log n)$-space deterministic index for S such that later, given a pattern $P[1..m]$, we can find all *occ* occurrences of P in S in $O(m \log m + occ \log \log n)$ time.

3.6 (Approximate) Pattern matching in LZW-compressed texts

Paweł Gawrychowski (MPI für Informatik – Saarbrücken, DE)

Pattern matching is the most basic problem concerning processing text data. Its complexity seems rather well-understood, and there are quite a few very efficient algorithms that can be used to solve it. What seems not that well-understood is its compressed variant, where instead of the text (or the pattern) we are given its compressed representation, and the goal is to achieve running time depending on the size of this representation and not the original length. I will present a number of results concerning LZW-compressed pattern matching, which is pattern matching for texts compressed using Lempel-Ziv-Welch-like methods. Such methods are simple to implement, so they are used in practice, and on the other hand they are complicated enough to be interesting from the theoretical point of view. It turns out that even when both the text and the pattern are LZW-compressed, it is possible to solve pattern matching in linear time, where linear means linear in the size of the compressed representation. I will present some ideas used to prove this. Of course in practice we are more interested in approximate pattern matching, meaning pattern matching with errors or with mismatches. The previous results for approximate LZW-compressed pattern matching were based on a more or less blackbox application of some uncompressed approximate pattern matching tools, and hence took at least $O(nm)$ time even when the bound on the number or mismatches/errors was very small. I will discuss a recent result with Straszak where we achieve $O(n\sqrt{m})$ for constant k.

3.7 Algorithms on grammar based strings

Shunsuke Inenaga (Kyushu University, JP)

Joint work of Inenaga, Shunsuke; Bannai, Hideo; Masayuki, Takeda; I, Tomohiro; Gawrychowski, Paweł; Shinohara, Ayumi; Narisawa, Kazuyuki; Gagie, Travis; Lewenstein, Moshe; Landau, Gad; Goto, Keisuke; Yamamoto, Takanori; Tanaka, Toshiya; Matsubara, Wataru

Straight-line programs (SLPs) are widely accepted abstract model of outputs of grammar-based text compression algorithms. In this survey talk, I introduce our recent results on algorithms that process given SLPs efficiently. Our methods allow various operations on SLPs, such as computing q-gram frequencies, finding palindromes, squares, runs, etc. All of our algorithms do not explicitly decompress given SLPs, and run in time polynomial in the input size.

3.8 Local recompression for compressed text

Artur Jeż (MPI für Informatik – Saarbrücken, DE)

Main reference A. Jeż, "Faster Fully Compressed Pattern Matching by Recompression," in Proc. of the 39th Int'l Colloquium on Automata, Languages, and Programming (ICALP'12), LNCS, Vol. 7391, pp. 533–544, Springer, 2012.
URL http://dx.doi.org/10.1007/978-3-642-31594-7_45
URL http://arxiv.org/abs/1111.3244

In this talk we present a simple and natural local recompression technique that is applicable to implicit representations of text, such as grammars, compressed representations or word equations. In essence the method aims at having all the strings in the instance compressed in the same way. The compression is achieved by two simple rewriting rules: pair compression that replaces all appearances of a pair of different letters ab with a fresh letter c and block compression which replaces maximal block of the form a^k with a fresh letter a_k, for all possible k. The crucial part of the method is that we can apply those rules directly to the implicit representation, however, in order to do this we may need to change this representation a bit. Our changes are local and boil down to replacement of a nonterminal (variable , piece of compressed data etc.) X with bX or Xa. With appropriate choice of pairs to compress it can be shown that the length of the strings in the instance drop by a constant factor in each phase and on the other hand the size of the instance remains linear in the input size. The method finds application in checking the equivalence of two SLPs, fully compressed pattern matching (for SLPs), word equations and approximation of the smallest grammar. In all those cases the algorithm based on recompression matches or improves the currently best known.

3.9 List Update for Data Compression

Alejandro López-Ortiz (University of Waterloo, CA)

Joint work of López-Ortiz, Alejandro; Kamali, Shahin; Ladra, Susana; Seco, Diego; Dorrigiv, Reza

From inception, list update (LU) has been used as a means to compress data. In this talk we review the main practical results on the use of list update algorithms for data compression. We discuss the theoretical foundations of these results, then we present an LU-based compressing scheme which is superior to BWT. Interestingly enough this compression-inspired strategy also proves superior to MTF in the MRM cost model of Martinez, Roura and Munro in practice.

3.10 Indexing Graphs for Path Queries with Applications in Genome Research

Veli Maekinen (University of Helsinki, FI)

Joint work of Sirén, Jouni; Vaelimaeki, Niko; Maekinen, Veli
Main reference J. Sirén, N. Välimäki, V. Mäkinen, "Indexing Finite Language Representation of Population Genotypes," in Proc. of the 11th Int'l Workshop on Algorithms in Bioinformatics, LNCS, Vol. 6833, pp. 270–281, Springer, 2011.
URL http://dx.doi.org/10.1007/978-3-642-23038-7_23

We propose a generic approach to replace the canonical sequence representation of genomes with graph representations, and study several applications of such extensions. The technical tool is to extend Burrows- Wheeler transform (BWT) of strings to acyclic directed labeled graphs, to support path queries as an extension to substring searching. We develop, apply, and tailor this technique to the following applications: a) read alignment on an extended BWT index of a graph representing reference genome and known variants of it; b) split-read alignment on an extended BWT index of a splicing graph; c) read alignment on an extended BWT index of a phylogenetic tree of partial-order graphs. Other possible applications include probe/primer design and alignments to assembly graphs. The main focus in this article is on a), for which several technical and compatibility issues had to be resolved to make the approach practical. For index construction we develop a space-efficient algorithm that scales to human genome data. For queries we extend an efficient search-space pruning technique to enable approximate searches. For compatability we tailor the approach so that alignments to paths can be projected back to the reference genome, so that the results can be seamlesly plugged inside widely adopted workflows for variation calling. Finally, we report several experiments on the feasibility of the approach to these applications. This talk is based on journal version (under preparation) of our work in WABI 2011 [1].

References

1 Jouni Sirén, Niko Välimäki, and Veli Mäkinen. *Indexing Finite Language Representation of Population Genotypes*. In Proc. WABI, LNCS 6833, pp. 270–281, Springer, 2011.

3.11 XML Compression via DAGs Unranked trees can be represented using their minimal dag (directed acyclic graph)

Sebastian Maneth (University of Oxford, UK)

Joint work of Markus Lohrey; Sebastian Maneth; Mireille Bousquet-Mélou; Eric Noeth
Main reference M. Lohrey, S. Maneth, E. Noeth, "XML compression via DAGs," in Proc. of the 16th Int'l Conf. on Database Theory (ICDT'13), pp. 69–80, ACM, 2013.
URL http://dx.doi.org/10.1145/2448496.2448506

For XML this achieves high compression ratios due to their repetitive mark up. Unranked trees are often represented through first child/next sibling (fcns) encoded binary trees. We study the difference in size (i.e., number of edges) of minimal dag versus minimal dag of the fcns encoded binary tree. One main finding is that the size of the dag of the binary tree can never be smaller than the square root of the size of the minimal dag, and that there are examples that match this bound. We introduce a new combined structure, the "hybrid dag", which is guaranteed to be smaller than (or equal in size to) both dags. Interestingly, we find through experiments that last child/previous sibling encodings are much better for XML compression via dags, than fcns encodings. This is because optional elements are more likely to appear towards the end of child sequences. The talk is based on an ICDT'2013 paper with title "XML Compression via DAGs" coauthored with Eric Noeth and Markus Lohrey.

At the end of the talk we present some new results about the expected sizes of unranked and binary DAGs. The new results can be found in the long version of the above ICDT paper, authored by Lohrey, Maneth, Noeth, and Mireille Bousquet-Mélou.

3.12 Succinct Data Structures

J. Ian Munro (University of Waterloo, CA)

Main reference J.I. Munro, S.S. Rao, "Succinct Representations of Data Structures," Chapter 37 in Mehta and Sahni (eds.), Handbook of Data Structures and Applications, Chapman & Hall/CRC, 2005.

In this talk we give and overview of the historical development of succinct data structures for the representation of graphs from the late 1980's to the present. Early work focused on trees and planar graphs in space roughly the information theoretic minimum while supporting an increasing array of navigation operations in constant time. A variety of tree representation protocols was developed to support these operations. Much, perhaps even most, of the work was motivated by applications to text indexing. Later work on trees led to some unification of the representation protocols. Other work led to succinct representations of combinatorial structures such as groups, functions, permutations and partial orders. These ideas fed back into applications to tree representations and graph representations in general. Other aspects dealt with lower bounds and time space tradeoffs.

3.13 Indexing Highly Repetitive Collections

Gonzalo Navarro (University of Chile, CL)

Main reference G. Navarro, "Indexing Highly Repetitive Collections," in Proc. of the 23rd Int'l Workshop on Combinatorial Algorithms (IWOCA'12), LNCS, Vol. 7643, pp. 274–279, Springer, 2012.
URL http://dx.doi.org/10.1007/978-3-642-35926-2_29

The need to index and search huge highly repetitive sequence collections is rapidly arising in various fields, including computational biology, software repositories, versioned collections, and others. In this talk we describe the progress made along three research lines to address the problem: compressed suffix arrays, grammar compressed indexes, and Lempel-Ziv compressed indexes. Those lines offer progressively better compression but less search efficiency, which raises the challenge of achieving the best in both aspects. Other extended problems, such as searching in a range of versions, document listing, searching for complex patterns, etc. are outlined at the end.

3.14 Categorical Range Reporting

Yakov Nekrich (University of Kansas, US)

In colored range reporting problem a set of colored points is stored in a data structure. For a query rectangle Q, we must enumerate all distinct colors of points in Q. In this talk we give an overview of previous and new results and techniques for this important problem.

3.15 How to Cook a Poset

Patrick K. Nicholson (University of Waterloo, CA)

Joint work of Munro, J. Ian; Nicholson, Patrick K.
Main reference J. Ian Munro, P.K. Nicholson, "Succinct Posets," in Proc. of the 20th Annual European Symp. on Algorithms (ESA'12), LNCS, Vol. 7501, pp. 743–754, Springer, 2012.
URL http://dx.doi.org/10.1007/978-3-642-33090-2_64

In this talk we survey data structures for representing partially ordered sets, or posets. The first part of the talk provides definitions of terminology related to posets. The second part is a survey of data structure results. Our main focus are the recent results of Farzan and Fischer (ISAAC 2011), and Munro and Nicholson (ESA 2012), which apply techniques from the area of succinct data structures.

3.16 Bioinformatics Algorithms: Sequence Analysis, Genome Rearrangements, and Phylogenetic Reconstruction

Enno Ohlebusch (Universität Ulm, DE)

Powerful new techniques have revolutionized the field of molecular biology. The vast amount of DNA sequence information produced by next-generation sequencers demands new bioinformatics algorithms to analyze the data.

This book provides an introduction to algorithms and data structures that operate efficiently on strings (especially those used to represent long DNA sequences). It focuses on algorithms for sequence analysis (string algorithms), but also covers genome rearrangement problems and phylogenetic reconstruction methods.

3.17 Wavelet Tries

Giuseppe Ottaviano (University of Pisa, IT)

Joint work of Grossi, Roberto; Ottaviano, Giuseppe
Main reference R. Grossi, G. Ottaviano, "The Wavelet Trie: Maintaining an Indexed Sequence of Strings in Compressed Space," arXiv:1204.3581v1 [cs.DS] , 2012.
URL http://arxiv.org/abs/1204.3581v1

We introduce and study the problem of compressed indexed sequence of strings, i.e. representing indexed sequences of strings in nearly-optimal compressed space, while preserving provably good performance for the supported operations. We present a new data structure for this problem, the Wavelet Trie, which combines the classical Patricia trie with the wavelet tree, a succinct data structure for storing compressed sequences. The resulting Wavelet Trie smoothly adapts to a sequence of strings that changes over time. It improves on the state-of-the-art compressed data structures by supporting a dynamic alphabet (i.e., the set of distinct strings) and prefix queries, both crucial requirements in the aforementioned applications, and on traditional indexes by reducing space occupancy to close to the entropy of the sequence.

3.18 Lightweight Lempel-Ziv Parsing

Simon J. Puglisi (University of Helsinki, FI)

Joint work of Puglisi, Simon J.; Kempa, Dominik; Kärkkäinen, Juha
Main reference J. Kärkkäinen, D. Kempa, S. J. Puglisi, "Lightweight Lempel-Ziv Parsing," in Proc. of the 12th Int'l Symp. on Experimental Algorithms (SEA'13), LNCS, Vol. 7933, pp. 139–150, Springer, 2013.
URL http://dx.doi.org/10.1007/978-3-642-38527-8_14

We introduce a new approach to LZ77 factorization that uses $O(n/d)$ words of working space and $O(dn)$ time for any $d \geq 1$ (for polylogarithmic alphabet sizes). We also describe carefully engineered implementations of alternative approaches to lightweight LZ77 factorization. Extensive experiments show that the new algorithm is superior, and particularly so at the lowest memory levels and for highly repetitive data. As a part of the algorithm, we describe new methods for computing matching statistics which may be of independent interest.

3.19 Encoding Top-k Queries

Rajeev Raman (University of Leicester, UK)

Main reference R. Grossi, J. Iacono, G. Navarro, R. Raman, S. Rao Ratti, "Encodings for Range Selection and Top-k Queries," in Proc. of the 21st Annual European Symp. on Algorithms (ESA'13), LNCS, Vol. 8125, pp. 553–564, Springer, 2013.
URL http://dx.doi.org/10.1007/978-3-642-40450-4_47

We study the problem of *encoding* the positions the top-k elements of an array $A[1..n]$ for a given parameter $1 \leq k \leq n$. Specifically, for any i and j, we wish create a data structure that reports the positions of the largest k elements in $A[i..j]$ in decreasing order, *without* accessing A at query time. This is a natural extension of the well-known encoding range-maxima query problem, where only the position of the maximum in $A[i..j]$ is sought, and finds applications in document retrieval and ranking. We give (sometimes tight) upper and lower bounds for this problem and some variants thereof for general k [1], and a solution for the specific case of $k = 2$ that has better constants than the general solution above [2].

References

1 Roberto Grossi, John Iacono, Gonzalo Navarro, Rajeev Raman and S. Srinivasa Rao. *Encodings for Range Selection and Top-k Queries*. In Proc. ESA, LNCS 8125, pp. 553-564, Springer, 2013.
2 Pooya Davoodi, Gonzalo Navarro, Rajeev Raman and S. Srinivasa Rao. *Encoding Range Minimum Queries*. Manuscript, 2013.

3.20 Compressed Pattern Matching on Terms

Manfred Schmidt-Schauss (Goethe-Universität Frankfurt am Main, DE)

Terms that are compressed with a singleton tree grammar are considered. The fully compressed pattern (sub-)match within a compressed term is analysed, trying to find the special cases which permit a polynomial time algorithm. Such cases are: (i) where compressed patterns contain every variable at most once; (ii) the pattern is DAG-compressed. It is open whether there exists a polynomial time algorithm for the general case.

3.21 Top-k Document Retrieval

Rahul Shah (Louisiana State University, US)

Joint work of Hon, Wing-Kai; Shah, Rahul; Thankachan, Sharma T.; Vitter, Jeffrey S.

Document retrieval is a special type of pattern matching that is closely related to information retrieval and web searching. In this problem, the data consist of a collection of text documents, and given a query pattern P, we are required to report all the documents (not all the occurrences) in which this pattern occurs. In addition, the notion of *relevance* is commonly applied to rank all the documents that satisfy the query, and only those documents

with the highest relevance are returned. Such a concept of relevance has been central in the effectiveness and usability of present day search engines like Google, Bing, Yahoo, or Ask. When relevance is considered, the query has an additional input parameter k, and the task is to report only the k documents with the highest relevance to P, instead of finding all the documents that contain P. For example, one such relevance function could be the frequency of the query pattern in the document. In the information retrieval literature, this task is best achieved by using inverted indexes. However, if the query consists of an arbitrary string—which can be a partial word, multiword phrase, or more generally any sequence of characters—we cannot take advantages of the word boundaries and we need a different approach. This leads to one of the active research topics in string matching and text indexing community in recent years, and various aspects of the problem have been studied, such as space-time tradeoffs, practical solutions, multipattern queries, and I/O-efficiency. In this talk, we review some of the initial frameworks for designing such indexes and also summarize more recent developments in this area.

3.22 Semi-local LCS: Superglue for string comparison

Alexander Tiskin (University of Warwick, UK)

The computation of a longest common subsequence (LCS) between two strings is a classical algorithmic problem. A generalisation of this problem, which we call semi-local LCS, asks for the LCS between a string and all substrings of another string, and/or the LCS between all prefixes of one string and all suffixes of another. This generalised problem turns out to be fundamental whenever a solution to a sting comparison or approximate matching problem has to be "glued together" from its solutions on substrings: for example, approximate matching in a compressed string; comparing strings in parallel; dynamic support of a string comparison score. The semi-local LCS problem has an elegant algebraic structure, expressed by the monoid of "seaweed braids" (i.e., the 0-Hecke monoid of the symmetric group). It also has surprising connections with computational geometry, planar graph algorithms, comparison networks, as well as practical applications in computational molecular biology. We discuss efficient algorithms for the semi-local LCS problem, and survey some related results and applications.

3.23 Distributed String Mining

Niko Vaelimaeki (University of Helsinki, FI)

Joint work of Vaelimaeki, Niko; Puglisi, Simon J.
Main reference N. Välimäki, S.J. Puglisi, "Distributed String Mining for High-Throughput Sequencing Data," in Proc. of the 12th Int'l Workshop on Algorithms in Bioinformatics (WABI'12), LNCS, Vol. 7534, pp. 441–452, Springer, 2012.
URL http://dx.doi.org/10.1007/978-3-642-33122-0_35

The goal of frequency constrained string mining is to extract substrings that discriminate two (or more) datasets. Known solutions to the problem range from an optimal time algorithm

to different time-space tradeoffs. However, all of the existing algorithms have been designed to be run in a sequential manner and require that the whole input fits the main memory. Due to these limitations, the existing algorithms are practical only up to a few gigabytes of input. We introduce a distributed algorithm that has a novel time-space tradeoff and, in practice, achieves a significant reduction in both memory and time compared to state-of-the-art methods. To demonstrate the feasibility of the new algorithm, our study includes comprehensive tests on large-scale metagenomics data. We also study the cost of renting the required infrastructure from, e.g. Amazon EC2. Our distributed algorithm is shown to be practical on terabyte-scale inputs and affordable on rented infrastructure.

References

1 Nieves R. Brisaboa, Rodrigo Canovas, Francisco Claude, Miguel A. Martinez-Prieto and Gonzalo Navarro. *Compressed String Dictionaries*. In: Proc. SEA, LNCS 6630, pp. 136–147. Springer, 2011.

2 Roberto Grossi and Giuseppe Ottaviano. *Fast Compressed Tries through Path Decompositions*. In: Proc. ALENEX, pp. 65–74, SIAM, 2012.

Participants

- Djamal Belazzougui
University of Helsinki, FI
- Philip Bille
Technical University of Denmark – Lyngby, DK
- Stefan Böttcher
Universität Paderborn, DE
- Francisco Claude
University of Waterloo, CA
- Henning Fernau
Universität Trier, DE
- Johannes Fischer
KIT – Karlsruhe Institute of Technology, DE
- Travis Gagie
University of Helsinki, FI
- Paweł Gawrychowski
MPI für Informatik – Saarbrücken, DE
- Roberto Grossi
University of Pisa, IT
- Shunsuke Inenaga
Kyushu University, JP
- Artur Jeż
MPI für Informatik – Saarbrücken, DE
- Juha Kärkkäinen
University of Helsinki, FI
- Susana Ladra Gonzalez
University of La Coruna, ES
- Alejandro Lopez-Ortiz
University of Waterloo, CA
- Veli Mäkinen
University of Helsinki, FI
- Sebastian Maneth
University of Oxford, GB
- J. Ian Munro
University of Waterloo, CA
- Gonzalo Navarro
University of Chile, CL
- Yakov Nekrich
Univ. of Kansas – Lawrence, US
- Patrick K. Nicholson
University of Waterloo, CA
- Enno Ohlebusch
Universität Ulm, DE
- Giuseppe Ottaviano
University of Pisa, IT
- Simon J. Puglisi
University of Helsinki, FI
- Rajeev Raman
University of Leicester, GB
- Manfred Schmidt-Schauss
Goethe-Universität Frankfurt am Main, DE
- Diego Seco
University of Conception, CL
- Rahul Shah
Louisiana State University, US
- Yasuo Tabei
Hokkaido University, JP
- Sharma V. Thankachan
Louisiana State University, US
- Alexander Tiskin
University of Warwick, GB
- Koji Tsuda
CBRC – Tokyo, JP
- Niko Välimäki
University of Helsinki, FI
- Rossano Venturini
University of Pisa, IT
- Oren Weimann
Haifa University, IL

Report from Dagstuhl Seminar 13241

Virtual Realities

Edited by

Guido Brunnett[1], Sabine Coquillart[2], Robert van Liere[3], and Gregory Welch[4]

1 **TU Chemnitz, DE,** guido.brunnett@informatik.tu-chemnitz.de
2 **INRIA Rhone-Alpes, St. Ismier, FR,** Sabine.Coquillart@inria.fr
3 **CWI – Amsterdam, NL,** robert.van.liere@cwi.nl
4 **The University of Central Florida – Orlando, US,** welch@ucf.edu

Abstract

This report documents the program and the outcomes of Dagstuhl Seminar 13241 "Virtual Realities". The main goal of the five day seminar was to bring together leading experts and promising young researchers to discuss current challenges and future directions in the field of virtual and augmented reality. The seminar was organized as series of individual presentations and seven working groups. Abstracts of the presentations and working group reports are collected in this report.

Seminar 09.–14. June, 2013 – www.dagstuhl.de/13241
1998 ACM Subject Classification I.3.7 [Three- Dimensional Graphics and Realism]: Virtual Reality, I.3.1 [Hardware Architecture]: Three-Dimensional Displays, H.5.2 [Information Interfaces and Presentation]: User Interfaces – Interaction Styles; Graphical user interfaces
Keywords and phrases Virtual Reality, 3D Interaction, Presence, Human Factors
Digital Object Identifier 10.4230/DagRep.3.6.38
Edited in cooperation with Libor Váša

1 Executive Summary

Guido Brunnett
Sabine Coquillart
Robert van Liere
Gregory Welch

Virtual Reality (VR) is a multidisciplinary area of research aimed at interactive human computer mediated simulations of artificial environments. An important aspect of VR-based systems is the stimulation of the human senses – usually sight, sound, and touch – such that a user feels a sense of presence in the virtual environment. Sometimes it is important to combine real and virtual objects in the same real or virtual environment. This approach is often referred to as Augmented Reality (AR), when virtual objects are integrated into a real environment. Research in VR and AR encompasses a wide range of fundamental topics, including: 3D interaction, presence, telepresence and tele-existence, VR modelling, multi-model systems, and human factors. Typical VR applications include simulation, training, scientific visualization, and entertainment, whereas typical AR applications include computer-aided manufacturing or maintenance, and computer-aided surgery or medicine.

The main goal of the seminar was to bring together leading international experts and promising young researchers to discuss current VR and AR challenges and future directions.

The organization built on the experiences from the previous seminar "Virtual Realities 2008". The format of the seminar included sessions with standard presentations as well as parallel breakout sessions devoted to "hot-topics" in VR and AR research. It was the desire of the participants of the seminar that sufficient time for plenary discussion and working groups was scheduled. Before the seminar, the organizers solicited topics for the working groups. During the first days of the seminar these working groups were formed and a schedule was created. Plenary sessions were also scheduled to allow the working groups to report and discuss their findings.

Eight plenary sessions of presentations were scheduled throughout the week. Each session usually consisted of three 15 minute presentations followed by a 45 minute moderated discussion. Abstracts of the presentations are collected in the next chapter. The Monday afternoon plenary sessions were devoted to the topics of Telepresence and Human Embodiment. Tuesday morning the topics Applications and Health/Wellbeing were presented. Wednesday morning was devoted to a session on Virtual Environments. The Thursday morning sessions were on Commercial/Buisiness aspects of VR and Authoring/ Content. The last session was devoted to Augmented Reality.

Seven working groups were created and parallel breakout sessions held throughout the week. Each working group reported their findings in plenary sessions. The following lists the titles of the working groups:

- Real Time Interactive Systems – Architecture Issues
- VR Current State and Challenges
- 3D User Interfaces
- Avatars in Virtual Reality
- Scientific Visualization and VR
- Characterising Interactions in Virtual (and/or Real) Environments
- Unconventional Mixed Environments

2 Table of Contents

3 Overview of Talks

3.1 Human-subject experiments in Virtual Reality and 3DUI

Carlos Andujar (UPC – BarcelonaTech, ES)

This talk is about the major peculiarities and difficulties we encounter when trying to validate research results in fields such as virtual reality (VR) and 3D user interfaces (3DUI). We review the steps in the empirical method and discuss a number of challenges when conducting human-subject experiments. These challenges include the number of independent variables to control to get useful findings, the within-subjects or between-subjects dilemma, hard-to-collect data, experimenter effects, ethical issues, and the lack of background in the community for proper statistical analysis and interpretation of the results. We show that experiments involving human-subjects hinder the adoption of traditional experimental principles (comparison, repeatability, reproducibility, justification and explanation) and propose some ideas to improve the reliability of findings in VR and 3DUI disciplines.

3.2 "Live" will never be the same

Wolfgang Broll (TU Ilmenau, DE)

If you would just ask an arbitrary person on the street what she thinks VR is, she would probably answer: the matrix or maybe the Holodeck. Another possibility to express this would be to say: what would make a virtual environment undistinguishable from real life. There is no easy and no immediate answer to this question. However, if we look at AR, it seems that we are already much closer to a situation, where it becomes impossible or at least very difficult for an individual to distinguish between real and artificial (virtual) content. Recent works in the area of Diminished Reality and AR with real lighting reveal that the remaining steps in this area might be much smaller than usually estimated. Since adding live virtual content is already well established in the area of broadcasting and due to a rapidly growing market for virtual product placement, providing the necessary driving force, perfectly integrated sophisticated (live) virtual content will probably quite soon become a standard element within movies and broadcastings. Thus "live" transmissions will contain additional unreal content, while other real elements are discarded – not recognizable for the observer. Combined with recent advances in see-through display technologies we should not be surprised to already see individually adapted environments undistinguishable from (pure) reality within a couple of years. This seems feasible as in contrast to VR most of the content observed will still be real and virtual content can be adapted to the real one much easier than creating an entire convincible artificial world. As with the matrix and the Holodeck, this raises the question, how such a development will influence our daily life with respect to communication, interaction, and the reception of our environment.

3.3 A Multi-Projector CAVE with Gesture Recognition

Pere Brunet (UPC – BarcelonaTech, ES)

Joint work of Brunet, Pere; Andujar, Carlos; Vinacua, Alvar
Main reference The full paper is under review

In this talk, we present a novel four wall multi-projector CAVE architecture which is powered by 40 off-the-shelf projectors controlled by 12 PCs. It operates in passive stereo, providing high brightness at 2000 x 2000 pixel resolution on each of the 4 walls. We have achieved high resolution while significantly reducing the cost and increasing versatility: the system works with any mix of a wide range of projector models that can be substituted at any moment for more modern or cheaper ones. The uniformity of the final image is achieved using a specially designed self-calibration software which adapts each of the 40 projectors and guarantees concordance and continuity. The main contributions of our approach are: (a) The design and construction of a passive stereo, four-wall CAVE system with commodity hardware. It is based on 40 off-the-shelf DLP projectors and 12 PCs. The CAVE design achieves higher resolution and brightness while significantly reducing the total cost. (b) The system is versatile: it works with any mix of a wide range of projector models that can be substituted at any moment for more modern or cheaper ones. (c) Uniformity of the final image is guaranteed by a specially designed self-calibration software which adapts each of the 40 projectors and guarantees concordance and continuity. Independent self-calibration of the different CAVE walls is sufficient in most of the cases. And (d), A gesture-based, ergonomic interaction paradigm, based on dynamically merging the information from two orthogonal kinect sensors, has been designed and implemented. Interaction is intuitive and cableless.

3.4 Authoring VR in the Post-WIMP Age: Challenges, Approaches, Solutions

Ralf Doerner (Hochschule RheinMain – Wiesbaden, DE)

Joint work of Doerner, Ralf; Gerken, Katharina; Frechenhaeuser, Sven; Luderschmidt, Johannes
Main reference K. Gerken, S. Frechenhaeuser, R. Doerner, J. Luderschmidt, "Authoring Support for Post-WIMP Applications," in Proc. of the 14th IFIP TC 13 Int'l Conf. on Human-Computer Interaction (INTERACT'13), LNCS, Vol. 8119, pp. 744–761, Springer, 2013.
URL http://dx.doi.org/10.1007/978-3-642-40477-1_51

Currently, VR is heavily influenced by the advent of Post-WIMP user interfaces that outstrip the WIMP (Windows, Icons, Menus, Pointer) – paradigm of traditional PC user interfaces. Because of a larger user base, robust and reasonably priced hardware (e.g. mobile tablet PCs with multitouch screen, Microsoft Kinect, Leap Motion, Google Glasses) becomes available that can be readily used in novel VR setups with an increasing number of sensors. Users become acquainted with more "exotic' input and output methods facilitating novel interaction designs. VR setups can be put in more contexts (e.g. living spaces) where users "own" parts of the VR setup. This allows for novel application areas (e.g. ambient intelligence), more complex VR content (e.g. virtual humans with sophisticated behavior) and introduces new authoring groups to VR (especially authors with non-technical background such as designers, domain experts or end users). One major limiting factor for exploiting this new potential

for VR is the problem that authoring processes are often not suited for new author groups and not able to deal with its complexity, requiring hand-crafted solutions which are costly to produce and to adapt. Several approaches to tackle problems associated with authoring aspects will be presented in the talk. First, novel authoring tools based on the Post-WIMP paradigm are introduced. Second, authoring processes relying on components and meta data are put forward. Third, the application of Self-X methodologies (e.g. Self-Organization or Self-Adaptation) to the field of VR is examined (e.g. creators of VR systems can be freed from tasks such as manual integration of additional sensors in a VR system and calibration if this can be performed automatically using Self-Configuration). Fourth, complex event processing (and according languages such as Esper) can be employed to provide VR application developers with more abstract events allowing for easy handling of large volumes of sensor data. The discussion of these approaches will cover ideas, solutions and visions for VR authoring. Also the question what VR can contribute to Post-WIMP paradigms will be addressed. The provision of suitable methodologies for authoring VR systems and creating VR content is a crucial issue for the success of VR since it determines key characteristics of VR applications such as quality, usability, applicability, flexibility and cost efficiency.

3.5 Three Topics in Designing Augmented Reality

Steven K. Feiner (Columbia University, US)

Joint work of Feiner, Steven K.; Dedual, Nick; Henderson, Steve; Oda, Ohan; Sukan, Mengu; Tversky, Barbara

As Augmented Reality (AR) technologies stand poised to become a part of our daily lives, how can we design effective user interfaces that incorporate them? I will provide brief overviews of research being done by Columbia's Computer Graphics and User Interfaces Lab on three topics:

- How can we use AR to assist users in training for and performing tasks, alone and in collaboration with others? What can we do to avoid providing too much information, while helping users understand the task well enough that they can be effective even if the AR technology were to fail.
- What is the place in AR for points of view other than the immediate first- person?
- How can we combine multiple heterogeneous displays in AR, to benefit from their complementary advantages?

References

1 Nicolas Dedual, Ohan Oda, and Steven Feiner. Creating hybrid user interfaces with a 2D multi-touch tabletop and a 3D see-through head-worn display. *Proc. ISMAR 2011* (*IEEE Int. Symp. on Mixed and Augmented Reality*), Basel, Switzerland, October 26–29, 2011, 231–232.
2 Nicolas Dedual and Steven Feiner. Addressing information overload in urban augmented reality applications. *Proc. GeoHCI 2013* (*CHI 2013 Workshop on Geographic Human–Computer Interaction*), Paris, France, April 27–28, 2013.
3 Steven Henderson and Steven Feiner. Augmented reality in the psychomotor phase of a procedural task. *Proc. ISMAR 2011* (*IEEE Int. Symp. on Mixed and Augmented Reality*), Basel, Switzerland, October 26–29, 2011, 191–200.

4 Ohan Oda, Mengu Sukan, Steven Feiner, and Barbara Tversky. Poster: 3D referencing for remote task assistance using augmented reality. *Proc. 3DUI 2013* (*IEEE Symp. on 3D User Interfaces*), Orlando, FL, March 16–17, 2013, 179–180.
5 Mengu Sukan, Steven Feiner, Barbara Tversky, and Semih Energin. Quick viewpoint switching for manipulating virtual objects in hand-held augmented reality using stored snapshots. *Proc. ISMAR 2012* (*IEEE Int. Symp. on Mixed and Augmented Reality*), Atlanta, GA, November 5–8, 2012, 217–226.

3.6 Local and Remote Collaboration in Multi-User Virtual Reality

Bernd Froehlich (Bauhaus-Universität Weimar, DE)

Joint work of Beck, Stephan; Kunert, André; Kulik, Alexander; Froehlich, Bernd
Main reference S. Beck, A. Kunert, A. Kulik, B. Froehlich, "Immersive Group-to-Group Telepresence," IEEE Trans. on Visualization and Computer Graphics, 19(4):616–25, March 2013 (Proceedings of IEEE Virtual Reality 2013, Orlando, Florida).
URL http://dx.doi.org/10.1109/TVCG.2013.33

Our immersive telepresence system [1] allows distributed groups of users to meet in a shared virtual 3D world. Our approach is based on two coupled projection-based multi-user setups, each providing multiple users with perspectively correct stereoscopic images [3]. At each site the users and their local interaction space are continuously captured using a cluster of registered depth and color cameras. The captured 3D information is transferred to the respective other location, where the remote participants are virtually reconstructed. Local and remote users can jointly or independently explore virtual environments and virtually meet face-to-face for discussions. We structure collaborative activities of collocated and remote users using Photoportals [2]. Virtual photos and videos serve as three-dimensional references to objects, places, moments in time and activities of users. They can be shared among users and serve as portals to the captured information. Our Photoportals also provide access to intermediate or alternative versions of a scenario and allow the review of recorded task sequences that include life-size representations of the captured users.

References

1 Beck, S., Kunert, A., Kulik, A., Froehlich, B. *Immersive Group-to-Group Telepresence*, IEEE Transactions on Visualization and Computer Graphics, 19(4):616-25, March 2013
2 Kunert, A., Kulik, A., Beck, S., Froehlich, B., *Photoportals: Shared References in Space and Time*, In proceedings of ACM CSCW, February 2014, Baltimore, to appear.
3 Kulik A., Kunert A., Beck S., Reichel R., Blach R., Zink A., Froehlich B. *C1x6: A Stereoscopic Six-User Display for Co-located Collaboration in Shared Virtual Environments*. ACM Transactions on Graphics 30, 6, Article 188 (December 2011), 12 pages.

3.7 Telepresence Systems: What Is To Be Done

Henry Fuchs (University of North Carolina at Chapel Hill, US)

Joint work of Maimone, Andrew; Yang, Xubo; Dierk, Nate; State, Andrei; Dou, Mingsong; Fuchs, Henry
Main reference A. Maimone, X. Yang, N. Dierk, A. State, M. Dou, H. Fuchs, "General-Purpose Telepresence with Head-Worn Optical See-Through Displays and Projector-Based Lighting," in Proc. of 2013 IEEE Virtual Reality (VR'13), pp. 23–26, IEEE, 2013.
URL http://dx.doi.org/10.1109/VR.2013.6549352
URL http://www.cs.unc.edu/ maimone/media/GeneralTelepresence_VR_2013.pdf

We have known for decades what is needed for an effective telepresence system: 3D acquisition of the remote scene, and 3D display of the remote scene to the local participant. For flexibility, the display of the remote scene should be an augmentation of the local participant's view of his or her local surroundings. Variations of telepresence systems often appear in science fiction films and more recently in television, even in news programs. All of these appearances to date are achieved by special effects of some sort, since the technology to achieve telepresence is still inadequate for actual use. The widespread availability of two different technology trends in the past few years have accelerated research interest in telepresence: 1) inexpensive depth cameras, most notably Microsoft Kinect, and 2) small high-resolution displays for mobile devices, from which head-worn displays can be built. Significant challenges still remain, however. For a useful telepresence system, a scene capture using multiple depth cameras is likely necessary. Unfortunately, the active (infra-red) light-based technology in Kinect depth cameras degrades significantly, due to interference, when multiple units are used in the same scene. The other part of the system, the display, is also still a challenge. Augmented visualization displayed within the local user's surroundings still cannot be achieved effectively. Such augmented visualization is desirably since in many, perhaps most situations, there are other local participants or local objects, with which the local participant wants to interact while interacting with the remote participant(s). To produce a proper combined augmented, mixed visualization, the system has to perform proper occlusion in 3D: the remote participant(s) have to occlude local objects behind them and they have to be occluded by local objects in front of them. Thus 3D scene acquisition is likely necessary for both local and remote spaces. In addition, some kind of optical or video see-through head-worn display may be needed for each of the participants. Current versions of such displays with occlusion capabilities are either very bulky or have narrow field of view. Until adequate such displays are developed, other alternatives may be more attractive: 1) large-format multi-viewer autostereo displays, and 2) simple optical see-through head-worn displays coupled with local light control accomplished with projector-based illumination. The availability of head-worn displays such as Google Glass and Lumus DK-32 will accelerate interest in and work on head-worn displays, which will eventually benefit telepresence applications and systems.

3.8 AR UI: Balancing Real and Virtual

Raphael Grasset (TU Graz, AT)

Augmented Reality has moved from lab environments to the hands of the general public over the last few years. Yet, designing Augmented Reality user interfaces (AR UI) for the

next generation of AR applications remains a general issue. Current commercial AR user interfaces mimics virtual reality or traditional user interface (e.g. touch screen) rather than fully capturing the essence of AR: Real and Virtual.

In this talk I give an overview of the limitations of current AR user interfaces and a list of case studies, based on my previous work ([4], [3], [1], [2]), on how we can better balance real and virtual in the design of AR UI. I describe how a seamless AR world can rely on better view management techniques, methods for creating virtual content out of real artefacts, and the effective combination of computer vision, computer graphics and HCI. Finally, I outline a research agenda and new principles for the development of future generation of AR UI: leveraging visual perception theories and methods, more adaptive and contextual UI, new development tools for UI design (simulator, datasets) and an AR Semiotics framework.

References

1 Raphael Grasset, Tobias Langlotz, Denis Kalkofen, Markus Tatzgern, and Dieter Schmalstieg. *Image-driven view management for augmented reality browsers*. ISMAR, page 177-186. IEEE Computer Society, (2012)

2 Tobias Langlotz, Mathäus Zingerle, Raphael Grasset, Hannes Kaufmann, Gerhard Reitmayr. *AR Record & Replay: Situated Compositing of Video Content in Mobile Augmented Reality*, Proceedings of OzCHI (Australian Computer-Human Interaction Conference), 2012.

3 Adrian, Clark, Andreas Dünser, Raphael Grasset, *An interactive augmented reality coloring book*. SIGGRAPH Asia 2011 Emerging Technologies, 2011.

4 Nate Hagbi, Raphael Grasset, Oriel Bergig, Mark Billinghurst, Jihad El-Sana, !In-Place Sketching for Content Authoring in Augmented Reality Games, Virtual Reality (VR'10), IEEE, 2010.

3.9 AR and VR Everywhere?

Tobias Hoellerer (University of California – Santa Barbara, US)

AR and VR hold enormous promises as paradigm-shifting ubiquitous technologies. We all have seen indications of this potential, but clearly most of us are still far from using AR or VR on a regular basis, everywhere. What will it take to get closer to mass adoption? My presentation discussed existing success stories and ongoing limitations of AR and VR technologies on the roadmap to seamless interaction with the physical world. New developments in sensors and real-time computer vision, coupled with near-to-universal connectivity may make it possible to finally scale the user interface experience from small screens to the wider context of the world before us. And on the way there, somebody will surely want your data...

3.10 Applications of Avatar Mediated Interaction to Teaching, Wellness & Education

Charles E. Hughes (University of Central Florida – Orlando, US)

Joint work of Hughes, Charles E.; Dieker, Lisa; Hynes, Michael; Nagendran, Arjun; Welch, Gregory

Teacher education usually involves practice with real children either as their sole teacher or in an apprentice role. This puts children (and teachers) at risk while skills, especially soft skills associated with human-to-human interaction, are still being learned. Virtual environments provide an opportunity for teachers to practice skills without interfering with the education of children and without placing themselves at the mercy of these same children while their classroom management, pedagogy and even content skills are still developing. Unfortunately, purely virtual worlds (ones driven by programmed behaviors) are not adaptive enough to provide realistic responses to verbal and non-verbal interactions, and are not capable of handling the almost random directions that the conversations may or should take. Similarly, virtual worlds that deal with protective strategies, e.g., for pre-teens facing enormous peer pressure, and those that deal with counselling situations, need to reflect the subtleties of human interaction that are not presently attainable through programmed behaviors, even those encompassing evolutionary changes in those behaviors.

The presentation focused on how a human-in-the-loop controlling the virtual avatars can provide the realism needed to address the areas mentioned above, as well as other applications involving intense human-to-human verbal and non- verbal interaction. Specifically, the talk focused on TeachLivE (Teach and Learning in a Virtual Environment), a system in current use at 32 universities and several school districts in the U.S. The TeachLivE system provides in-service and pre- service teachers the opportunity to practice skills and reflect on their own performances. The reflection component is achieved through an integrated after- action review system that support real-time and off-line tagging of events, along with the use of these tags to select video sequences that demonstrate the teacher's skills (or lack thereof) in different contexts. The system's scalability is enhanced by its micropose-based network protocol and its use of just one human inhabitor to control multiple avatars.

In addition to the applications already mentioned, the presentation discussed the uses of avatar-mediated interaction for children and young adults with autism and the employment of the underlying virtual settings in free-choice education.

References

1 Nagendran A, Pillat R, Kavanaugh A, Welch G, Hughes CE. (2013). AMITIES: Avatar-Mediated Interactive Training and Individualized Experiences System. Virtual Reality Software & Technology (VRST) 2013, Singapore, October 6-8, 2013

3.11 VELOS – A VR environment for ship applications: current status and planned extensions

Panagiotis D. Kaklis (National Technical University of Athens, GR)

VELOS (Virtual Environment for Life On Ships) is a multi-user VR system enabling passenger- and crew-activities for normal/hectic conditions. Ship evacuation currently constitutes its main application area. In the future we plan to extend its coverage with crew-ergonomics and -training as well as passenger comfortability. VELOS is based on VRsystem, a generic multi-user environment that adopts a client-server architecture and offers a variety of functionality including: geometric, topological and VR modeling, crowd-microscopic modeling, interfacing to simulation packages (so far: seakeeping behavior, fire evolution) and networking support.

In this presentation we focused on two higher-order steering behaviors, namely "passenger grouping" and "crew assistance", that we have recently embedded in VELOS by combining and/or enriching standard steering behaviors already available in VELOS. The performance of these behaviors has been illustrated through two tests. The first test was a generic one, involving the moving of 70 persons through a simple four-room configuration. The second test was associated with the evacuation of a ro-ro passenger ship, involving the movement of 100 passengers from their cabins in the after-zone of Deck 5 to the corresponding muster station on Deck 7. The recorded outcomes for both tests indicate that VELOS is capable for materializing complex evacuation scenarios for evaluating/modifying exiting general-arrangement layouts as well as crew training in immersive environments.

References

1 K.V. Kostas, A.-A.I. Ginnis, C.G. Politis, and P.D. Kaklis, *Motions effect for crowd modeling aboard ships*, short paper (poster exhibition) in Proceedings of the 6th International Conference on Pedestrian and Evacuation Dynamics, 5–8 June 2012, ETH, Zuerich (2012).

2 K.V. Kostas, A.I. Ginnis, C.G. Politis and P.D. Kaklis, *Use of VELOS platform for modeling and assessing crew assistance and passenger grouping in ship-evacuation analysis*, in the Proceedings of the Conference of the International Maritime Association of Mediterranean IMAM 2011, 13–16 September 2011, Genoa, Italy (2011).

3 A.I. Ginnis, K.V. Kostas, C. G. Politis and P.D. Kaklis, *VELOS: A VR Platform for Ship Evacuation Analysis*, CAD, 42, 1045–1058, (2010).

4 K.V. Kostas, A.-A.-A.I. Ginnis and P.D. Kaklis, *VELOS: A Virtual Environment for Life On Ships*, in Proceedings of the 3rd International Maritime Conference on Design for Safety (DFS2007), September 26–28, 2007, Berkeley, California, pp. 139–150.

5 K.V. Kostas, A.-A.I. Ginnis, P.D. Kaklis and A.D. Papanikolaou, *A VR-Environment for Investigating Passenger's Locomotion under Dynamic Ship Motion Conditions*, in Proceedings of the 8th International Marine Design Conference (IMDC-2003) , May 5–8, 2003, Athens, Greece, A.D. Papanikolaou (ed.), pp. 551–559.

3.12 VR Interfaces and Animation Algorithms for Modeling Autonomous Demonstrators

Marcelo Kallmann (University of California – Merced, US)

Joint work of Kallmann, Marcelo; Camporesi, Carlo; Mahmudi, Mentar; Huang, Yazhou
URL http://graphics.ucmerced.edu/index.html

This talk addresses the approach of developing animation algorithms and VR interfaces that are suitable for modeling motions to be used by autonomous demonstrators. VR interfaces allowing users to build motion clusters designed for inverse blending are presented, and planning methods on blending spaces are introduced for achieving humanlike variations that address obstacle avoidance for generic actions. The approach is complemented with locomotion synthesis in order to well position the virtual character for execution of given upper-body actions. As a result, autonomous virtual characters are able to plan and execute motions that are similar to demonstrated examples and that address new constraints and parameterizations.

3.13 Interactive Content for Well-being

Yoshifumi Kitamura (Tohoku University, JP)

Well-being is a concept which is used for a comprehensive understanding of subjective well-being, life satisfaction, positive emotion/affect, and so on, in positive psychology. Although it is difficult to achieve the well-being directly using the information technologies, I would like to conduct and accumulate my small pieces of effort toward the well-being by the research of interactive content. For this purpose, I have just started a research project to make a space of interpersonal interaction more active, enjoyable, efficient, comfortable, ... by research on interactive content. I introduced three research examples:

1. Understanding the "atmosphere" by verbal/nonverbal behaviors of persons measured by sensors,
2. D-FLIP: Dynamic & Flexible Interactive PhotoShow
3. TransformTable: A Self-Actuated Shape-Changing Digital Table

3.14 AR boundaries: human-factor implications of pro-longed HWD usage

Ernst Kruijff (Hochschule Bonn-Rhein-Sieg, DE)

Main reference E. Kruijff, J.E. Swan II, S. Feiner, "Perceptual issues in augmented reality revisited," in Proc. of the 9th IEEE Int'l Symp. on Mixed and Augmented Reality (ISMAR'10), pp. 3–12, IEEE, 2010.
URL http://dx.doi.org/10.1109/ISMAR.2010.5643530

Augmented reality (AR) is a field of research that has seen an incline in attention over the last years, driven foremost by simple cell phone applications. Nonetheless, the field of

research itself has steadily been growing over well over a decade. Predominantly, the focus has been on improving technical aspects of Augmented Reality, such as handling tracking or improving computing at small-scale platforms. As an effect of these efforts, systems have evolved from bulky installations towards powerful wearable platforms.

However, whereas AR itself is highly concerned with vision aspects, relatively little work has been performed on perception or cognition. As a result, many AR systems are suffering from the effects caused by problems such as depth distortion, focal attention switching, incorrect creation of object relationships, low legibility and occlusion. These issues limit the usability of AR applications considerably, in particular those systems that are based on so-called optical see- through systems. At the same time, both researchers and industry are in need of background information or even guidelines that may drive in particular applications that will be used by users in a pro-longed way. However, pro-longed usage is badly understood and may even further stress particular issues that already limit occasionally used applications. The presentation introduced several issues that merit further research, including visual tunneling, training effects, the usage of multiple perspectives (perceptual issues), and the effects on the user's cognitive map, search behavior, cognitive load, and mood alteration (cognitive or cognition- related issues). It is concluded that understanding long-term effects is important for sustaining interest in AR by both research and industry, as well as the actual end-users. To drive this understanding it will be necessary to lay out a structural approach to subsequently study and classify issues and effects.

3.15 Brain-Computer Interfaces and Virtual Environments

Anatole L"ecuyer (Inria, Rennes, FR)

Brain-Computer Interfaces (BCI) are introducing a novel user input for interaction with Virtual Environments: the user mental activity or cognitive processes. This input can be used in different ways, either for a direct control of virtual environments with "mental commands", or in an implicit interaction scheme by monitoring the cerebral activity and adapting the interaction or the content of the VE to the user "state of mind". Impressive prototypes exist today that combine immersive virtual reality technologies and BCI. But a lot a difficult scientific challenges remain for making this a robust and effective solution, involving multi-disciplinary research in Neuroscience and Neurophysiology, Signal-Processing, Human-Computer Interaction and Virtual Reality. But we believe we should see in the future promising simulators based on BCI and VR for industrial or medical training, rehabilitation and reeducation, or entertainment and videogames.

3.16 Lessons learned from introducing VR/AR in the German maritime industry

Uwe Freiherr von Lukas (FhG IGD, Rostock, DE)

In my presentation I shared several experiences we made at Fraunhofer IGD when introducing VR and AR to several areas of the maritime industry. There are many chances to use advanced visualization and interaction in the different stages of the lifecycle of a ship from design our production to training, operation and retrofit.

Due to the specific requirements of the maritime industry (one of a kind product, large data volume, projects under extreme time pressure) available standard VR solutions of other industries cannot be used off the shelf. So we recommend a use case-driven approach that leads to tailor-made solutions in terms of data integration, hardware setup, interaction and functionality. However, the efficiency in implementing those solutions and the return on investment of applying VR/AR in this way is still an open question.

3.17 Is there a significant market for industrial Virtual Reality today or tomorrow?

Uwe Freiherr von Lukas (FhG IGD, Rostock, DE)

The talk is mainly based on a study that characterizes and forecasts the German market for industrial 3D applications. It shows, that the market is dominated by small and medium-sized companies and that there are more service provider than hard- or software companies. The study states a market growth between 3% and 15% – depending on the penetration of 3D in new industrial processes. This growth is based on 3D technology spreading in various industries and processes. Classical VR does not have a significant share of the cake. This can be explained by the enormous complexity that VR systems still have and the necessity for well-trained staff to operate such an installation.

3.18 Perception & Action in Virtual Environments

Betty Mohler (MPI für biologische Kybernetik – Tübingen, DE)

URL http://www.youtube.com/user/MPIVideosProject

In the Perception and Action in Virtual Environments research group, our aim is to investigate human behavior, perception and cognition using ecologically valid and immersive virtual environments. Virtual reality (VR) equipment enables our scientists to provide sensory stimulus in a controlled virtual world and to manipulate or alter sensory input that would not be possible in the real world. More specifically, VR technology enables us to specifically manipulate the visual body, the contents of the virtual world, and the sensory stimulus (visual,

vestibular, kinesthetic, tactile, and auditory) while performing or viewing an action. Our group focuses on several different areas, all areas involve measuring human performance in complex everyday tasks, i.e. spatial judgments, walking, driving, communicating and spatial navigation. We investigate the impact of having an animated self-avatar on spatial perception, the feeling of embodiment or agency, and on the ability for two people to effectively communicate. Our goal is to use state-of-the-art virtual reality technology to better understand how humans perceive sensory information and act in the surrounding world. We use HMDs, large screen displays, motion simulators and sophisticated treadmills in combination with real-time rendering and control software and tools in order to immerse our participants in a virtual world. In this talk I will show videos of our technical setups and explain how I became interested in spatial perception in virtual reality. See specifically the following videos which are available on-line: http://www.youtube.com/user/MPIVideosProject.

References

1 Slater M, Spanlang B, Sanchez-Vives MV, Blanke O (2010) First Person Experience of Body Transfer in Virtual Reality. PLoS ONE 5(5): e10564. doi:10.1371/journal.pone.0010564

3.19 Human Body/Embodiment and Related Perceptions

Tabitha C. Peck (Duke University, US)

Joint work of Peck, Tabitha C.; Seinfeld, S; Aglioti, M; Slater, M
Main reference T.C. Peck, S. Seinfeld, S.M. Aglioti, M. Slater, "Putting Yourself in the Skin of a Black Avatar Reduces Implicit Racial Bias," Consciousness and Cognition, Vol. 22, Issue 3, pp. 779–787, September 2013.
URL http://dx.doi.org/10.1016/j.concog.2013.04.016

In this talk I present previous work that shows that it is possible to generate in people the illusory sense of ownership of a virtual body in immersive virtual reality. This can be achieved through synchronous multisensory stimulation with respect to the real and virtual body. However, the consequences of such embodiment have not been explored, for example, would embodiment of someone with racial bias in a body of the other racial group diminish their bias? Previous research suggested that physically embodying people in racially-different avatars decreased empathy and increased implicit racial bias. In this talk, I present results that demonstrate that embodiment in differently-raced avatars is possible and that embodying Caucasian participants in dark-skinned avatars for only ten minutes can significantly reduced implicit racial bias compared to embodying participants in light-skinned avatars. The results suggest that virtual environments and virtual embodiment seem promising for allowing the possibility to change socially negative attitudes and behaviour.

3.20 White Paper on Haptics

Jerome Perret (Haption – Aachen, DE)

Main reference Haptic-SIG, "White Paper on Haptics as a Contribution to the Horizon 2020 Framework Program," EUROVR, 2012.
URL http://www.hapticsig.org/sites/default/files/White%20paper%20on%20Haptics-final.pdf

In this talk, we outline the main scientific and technological challenges identified in the White Paper published by the Special Interest Group (SIG) on Haptics of the EuroVR Association. We focus on three main topics: haptic rendering, haptic technology, and standarization. In the domain of haptic rendering, further research is needed in order to enable the physical simulation of very large scenes including complex deformable objects, while keeping a high-frequency framerate. Regarding haptic technology, besides the growing demand for haptic/tactile feedback on mobile terminals, further effort should be invested in the development of low-cost and universal devices (including Open Source hardware), and software for haptic collaboration on networks. Standardization, supported by Open Source software libraries and middleware, will ensure that R&D efforts benefit all users.

3.21 VR for Disabled Persons: Current Research and Future Challenges

John Quarles (University of Texas at San Antonio, US)

Joint work of Quarles, John; Guo, Rongkai; Samaweera, Gayani
Main reference G. Samaraweera, R. Guo, J. Quarles, "Latency and Avatars in Virtual Environments and the Effects on Gait for persons with Mobility Impairments," in Proc. of the IEEE Symp. on 3D User Interfaces (3DUI'13), pp. 23–30, IEEE, 2013.
URL http://dx.doi.org/10.1109/3DUI.2013.6550192

The objective of this talk is to give insight into how to design effective virtual environments for disabled persons. Current knowledge of how disabled persons experience virtual reality is largely based upon virtual rehabilitation literature, which is very application specific. Moreover, almost all prior basic research has been performed with healthy participants. Thus, there is a significant need for basic, generalizable research results that evaluate disabled persons interaction and experience in virtual environments. This talk presents the latest research toward this goal, specifically summarizing findings of how disabled persons respond to latency, avatars, and other immersive stimuli can affect presence and interaction in virtual environments. Future challenges are discussed and future work is proposed. Ultimately, there is a need for much more basic research on how disabled people experience virtual reality.

3.22 Taking Augmented Reality out of the Laboratory and into the Real World

Christian Sandor (University of South Australia, AU)

URL http://www.magicvisionlab.com/

This presentation introduces our efforts to create commercial applications with Augmented Reality (AR), a user interface technology that overlays computer graphics over the user's view of their surroundings. Together with our industry partners, we investigate two scenarios: first, to mobile information browsing on mobile phones (Nokia); second, to industrial product design (Canon).

During the last decade, mobile information browsing on mobile phones has become a widely-adopted practice. This was made possible by the increase of wireless networking infrastructure and the ever increasing amount of online data. By employing AR, we enable users to access digital data much more fluidly and therefore assist everyday tasks much more effectively than with previous user interfaces. An example is AR X-Ray vision, which enables users to look through buildings and other obstacles.

A common task in industrial product design is to create physical prototypes of new products. A common prototyping method is to use 3D printers to create physical models. However, 3D printers are slow and expensive and changes to the shape are costly and labour intensive. We are investigating completely virtual prototypes that can be seen through AR and touched through a haptic device. This enables users to interactively change the shape and appearance of a prototype. After having successfully demonstrated virtual prototypes using a pen-shaped haptic device, we are currently developing a system that enables users to touch the virtual prototypes with all their fingertips.

Videos of our prototypes can be viewed at: http://www.magicvisionlab.com.

3.23 Augmented Reality Visualization Pipeline

Dieter Schmalstieg (TU Graz, AT)

Augmented Reality (AR) is a new medium and is primarily used for presenting visual information. Like with desktop computing, visualization techniques are needed to make sure human users understand the information presented in AR. However, principles of visualization are hardly applied in AR. This talk proposes to adapt a standard visualization pipeline for the needs of AR and organize the presentation of visual information in AR around the concepts of data transformation, filtering and rendering.

3.24 Augmented Reality in Altered Gravity

Oliver Staadt (Universität Rostock, DE)

Joint work of Markov-Vetter, Daniela; Mittag, Uwe; Staadt, Oliver
Main reference D. Markov-Vetter, E. Moll, O. Staadt, "Evaluation of 3D selection tasks in parabolic flight conditions: pointing task in augmented reality user interfaces," in Proc. of the 11th ACM SIGGRAPH Int'l Conf. on Virtual-Reality Continuum and its Applications in Industry (VRCAI'12), pp. 287–294, ACM, 2012.
URL http://dx.doi.org/10.1145/2407516.2407583

Intra-vehicular control and experiment support in manned spacecraft, such as the ISS, can benefit from Augmented Reality systems. To understand AR interaction under micro- and hyper-gravity conditions, we conducted a series of experiments during the 56th and 58th ESA Parabolic Flight Campaigns. We measured user performance for aimed pointing tasks during different phases of a parabolic flight: (i) normal gravity (1-g), (ii) zero gravity (0-g), and (iii) hyper gravity (1.8-g). In addition to pointing accuracy and speed, we obtained biofeedback of subjects through heart-rate variability (HRV) measurements.

We believe that the results of our experiments will influence the design of future Augmented Reality systems in altered-gravity environments.

References

1 Daniela Markov-Vetter, Anke Lehmann, Oliver G. Staadt, and Uwe Mittag, *Future interface technologies for manned space missions*, 62nd International Astronautical Congress, October 2011.
2 Daniela Markov-Vetter, Eckard Moll, and Oliver Staadt, *Evaluation of 3D selection tasks in parabolic flight conditions: pointing task in augmented reality user interfaces*, Proceedings of the 11th ACM SIGGRAPH International Conference on Virtual-Reality Continuum and its Applications in Industry (New York, NY, USA), VRCAI '12, ACM, December 2012, pp. 287–294.
3 Daniela Markov-Vetter, Eckard Moll, and Oliver G. Staadt, *Verifying sensorimotoric coordination of augmented reality selection under hyper- and microgravity*, International Journal of Advanced Computer Science **3** (2013), no. 5.

3.25 The Role of the Body in Perceiving Real and Virtual Spaces

Jeanine Stefanucci; Michael Geuss; Kyle Gagnon; Sarah Creem-Regehr

Our work investigates the perception of the body and space in real and virtual environments with the aim of determining whether observers view virtual environments as intended by designers. Using measures adopted from embodied perception theories in psychology, which emphasize the role of the body in space perception, we test whether observers perceive virtual spaces akin to real spaces in the context of body capabilities for action. In immersive virtual environments (IVEs) and real environments, we changed either the physical (real world) or virtual body to assess its influence on whether or not people said they could pass through or under an aperture. IVEs allowed for body manipulations that were not possible in the real world. We found that when the body was made wider or taller through physical manipulations in the real world, people's estimates of passing through or under an aperture

were altered along with their judgments of the width or height of the aperture. We also found that judgments of the ability to pass under or through an aperture were similar across real and virtual environments even when no changes to the body were implemented. Finally, we showed that virtual manipulations of body dimensions (some not possible in the real world) affected decisions about action with respect to apertures in IVEs. Overall, the findings suggest that the body plays a role in space perception in both real and virtual environments, suggesting that care should be taken when constructing virtual representations of the body, especially in the case of self avatars.

3.26 Telexistence

Susumu Tachi (Kaio University, JP)

Main reference S. Tachi, "Telexistence," World Scientific, ISBN-13 978-981-283-633-5, 2010.
URL http://tachilab.org/modules/publications/

3.26.1 What is Telexistence

Telexistence is a fundamental concept that refers to the general technology that allows a human being to experience a real-time sensation of being in a place other than his/her actual location and to interact with the remote environment, which may be real, virtual, or a combination of both [1]. It also refers to an advanced type of teleoperation system that allows an operator at the controls to perform remote tasks dexterously with the feeling of being in a surrogate robot working in a remote environment. Telexistence in the real environment through a virtual environment is also possible. Sutherland [2] proposed the first head-mounted display system, which led to the birth of virtual reality in the late 1980s. This was the same concept as telexistence in computer-generated virtual environments. However, it did not include the concept of telexistence in real remote environments. The concept of providing an operator with a natural sensation of existence in order to facilitate dexterous remote robotic manipulation tasks was called "telepresence" by Minsky [3] and "telexistence" by Tachi [4]. Telepresence and telexistence are very similar concepts proposed independently in the USA and in Japan, respectively. However, telepresence does not include telexistence in virtual environments or telexistence in a real environment through a virtual environment.

In this talk an overview is given of the telexistence manipulation system TELESAR up to version TELESAR V.

References

1 S. Tachi: Telexistence, World Scientific, ISBN-13 978-981-283-633-5, 2009.
2 I. E. Sutherland: A Head-Mounted Three Dimensional Display, Proceedings of the Fall Joint Computer Conference, pp.757-764, 1968.
3 M. Minsky: Telepresence, Omni, vol.2, no.9, pp.44-52, 1980.
4 S. Tachi, K. Tanie, and K. Komoriya: Evaluation Apparatus of Mobility Aids for the Blind, Japanese Patent 1462696, filed on December 26, 1980; An Operation Method of Manipulators with Functions of Sensory Information Display, Japanese Patent 1458263, filed on January 11, 1981.

3.27 Embodiment via Physical-Virtual Avatars

Gregory F. Welch (University of Central Florida – Orlando, US)

Joint work of Welch, Gregory F.; Hughes, Charles E.; Nagendran, Arjun; Bailenson, Jeremy N.; Slater, Mel

We have developed and demonstrated a general-purpose human surrogate system we call a Physical-Virtual Avatar (PVA). We can capture 2D imagery, head/facial motion, and audio of a moving, talking real human inhabiter, and reproduce those signals on the PVA, delivering a dynamic, real-time representation of the user to multiple viewers; while simultaneously presenting the inhabiter with dynamic imagery and audio of the remote participants and environment. Compared to conventional televideo or telepresence technologies we believe such a PVA system could be more useful in one-to-one or small group social situations where voice, eye gaze, facial expressions, body movement (kinesics), and in particular mobility and the use of space (proxemics) are more likely to affect the thoughts, feelings, and behaviors of the humans involved.

We are planning research and development aimed at improving our PVA methods and prototype systems, and using those systems to explore behaviors and beliefs about the remote existence of the human inhabiters. With our education and medical collaborators, we are planning to explore the use of PVAs to improve the quality of student access to peers and experts, and the quality of life for those with mobility restrictions such as hospital/home-bound individuals.

Our objectives include developing the graphical/visual, aural, computational, and system affordances of the PVA necessary (a) to create the illusion for an inhabiter that they are at a remote location, and (b) for remote participants that the inhabiter is with them; exploring the thoughts, feelings, and behaviors of the humans involved in PVA-mediated social interaction, and how they are influenced by the perceived presence of their counterparts; investigating the potential impact in application-driven scenarios involving education (e.g., cooperative learning and same age peer tutoring) and well-being (e.g., the psychological "escape" of humans confined to home or hospital); and fostering a community of researchers engaged in avatars and related social psychology.

4 Working Groups

4.1 Realtime Interactive Systems – Architecture Issues

Roland Blach (FhG IAO, Stuttgart, DE)

Realtime Interactive Systems (RIS) are an increasingly important field of research. Application areas range from Virtual, Mixed, and Augmented Reality (VR, MR, and AR) to advanced Human-Computer Interaction (HCI), realtime simulation, and computer games. Several RIS aspects are equally relevant to ambient and pervasive computing as well as to robotics. All these fields differ in many aspects, from concepts to method sets, and hence have their own research communities. But they additionally exhibit a very interesting and important intersection area when it comes to the software engineering parts and principles involved. The challenge of RIS systems-engineering lies in the contradictory coupling requirements:

tight coupling of system and user to maximize experience/immersion vs. loose coupling of components to master complexity.

Based on a proposal for the classification of RIS rooted in the VR/AR/MR domain – a RIS architecture questionnaire to capture individual system characteristics – the intention of the discussion was to get closer to an understanding of how to analyze and compare RIS from an architectural point of view instead of a feature based view. The proposed questionnaire was considered too large and the meaning of the questions is too fuzzy. Besides the questionnaire a thorough requirement analysis could be carried out but is probably too broad as the application domains of RIS are too heterogeneous. Another approach could be the analysis of 3-5 very different systems. Open questions are then: a) which criteria and b) which systems.

To give advice to system architects, another option could be to gather structured knowledge of adjacent domains as e.g. operating systems, X11, HTML5 or transactional memory.

Two contradicting trends for the generalization of RIS have been discussed. Many interaction researchers seem to be happy with commercially available systems which provide good development support and lively communities. This could lead to a defacto standard which also supports interoperability. Nevertheless it is important that these systems provide enough information of their internal behavior and timing that user studies can be interpreted meaningfully. On the other hand interaction researchers who work in areas where fast render times and low latency is necessary do not believe in general engines but want detailed control of the algorithms. This often leads to individual and heavily customized systems e.g. in case of tight coupling renderer and interaction. Also different graphical data necessitates other interaction paradigms which could possibly not be provided by standard system architectures.

All discussed points have added new information or confirmed existing knowledge to the state of the art, but no final conclusions can be drawn. The discussion will continue and the remarks and result of the breakout session will be fed back to the SEARIS (Software Engineering and Architectures for Realtime Interactive Systems) community and discussed further on the annual workshops. We hope that these discussions might improve the understanding of system patterns and consequently will lead to a more deliberate if not better system design. The discussion should be understood as a starting point for further discussions which the participants would be happy to continue.

4.2 VR Current State and Challenges

Carolina Cruz-Neira (University of Louisiana at Lafayette, US)

The members of this group were a well balanced variety of pioneers, expert researchers, industry practitioners, and young investigators. The conversation was guided by a series of discussion questions that were answered or, in some cases, debated, among the group's members. The group started the discussion answering the questions of what VR is today and what we think is still (if anything) exciting about VR. The consensus was that VR has become more pervasive and has infiltrated other fields, and therefore is not a "one size fits all" definition. The main excitement of VR is still the potential to provide experiences like no other technology and with the advances on the underlying technologies, the experiences are becoming increasingly more compelling and with higher fidelity. The group also discussed

the current barriers that are preventing a wider acceptance of VR and agreed that the most challenging issue was the lack of a common infrastructure to provide homogeneity across platforms, interactive devices and applications. The discussion continued with reviewing the current markets for VR, many of them already outside research labs and an overview of success stories. The discussion then evolved towards the challenges still left to address and the potential new opportunities that will be opened and enabled by addressing those challenges. The discussion wrapped up with the enthusiastic conclusion that VR is an exciting field, still very relevant, and still a great deal of fascinating challenges to solve.

4.3 3D User Interfaces

Rob Lindemann (Worcester, Massachusetts, US)

Since the previous Virtual Realities seminar (08231) at Dagstuhl in June 2008, there has been powerful movement towards a Democratization of 3D User Interaction, driven mainly by the reduction in cost of motion-sensing equipment, the wide-spread use of powerful mobile devices, and renewed interest by the general public in virtual reality through games.

The introduction of the Microsft Kinect in 2010 helped feed general interest in motion sensing for games started by the Nintendo Wii controller in 2006. Participants in the breakout session viewed this development as significant for both technical and social reasons. It created an intense sub-culture of "citizen researchers" who began experimenting with the technology, devising interesting and novel ways of using depth-camera input. Unbeknownst to these new 3DUI developers, many of the difficulties encountered in their works had been highlighted during the work of our community over the past two decades of 3DUI research. Despite this general lack of building on previous research, we felt it was positive that many new minds had been exposed to 3DUI development in general.

The global adoption of mobile devices, mainly iOS and Android, has created new interaction scenarios for on-the-go users. Mobile devices today typically provide several novel ways for users to interact with them, including accelerometers, gyroscopes, GPS, multi-touch surfaces, high-resolution displays, proximity sensors, voice input, and high-bandwidth connectivity. Social network usage has also led to new actions common across multiple applications, such as sharing, "liking," and location-based services, such as recommender systems and mapping. In addition, the use of Augmented Reality (AR) techniques has also seen increased interest in mobile settings, giving rise to new user interaction research problems to work on.

Many of the new interface devices have been born from a resurgence in video gaming, though wide-spread use of novel interface devices still seems elusive. Like the Wii before it, many players who were so excited to "get off the couch" to play in new and interesting ways have returned to their sedentary play styles; The hype around new devices seems to fade once players understand how much effort it takes to play. We feel this helps to highlight the need for continued, careful research in the field of 3DUI.

We also discussed what interesting research areas have appeared recently. One important one was a re-evaluation of the list of common tasks performed in immersive environments forwarded by Bowman et al. [1], in their seminal book on the topic. While there was widespread agreement that the common list of tasks, namely, Selection, Manipulation,

Navigation, Symbolic Input, and System Control, has served us well, some modifications should be proposed. It was felt (and supported by Ernst Kruijff, who was present at the session) that the System Control area was somewhat of a "grab bag" of things that didn't fit neatly into the other areas, and could be re-explored. In addition, there was some support for combining Selection and Manipulation into a single area, as it is often the case that users use them together in a single action, and separating them does not address the need to make the techniques used in a given system complimentary. This brings up another common feeling of the group, that there is still little work on selecting solutions for each of the tasks that work well together, and that reduce cognitive and technical load of switching between them. For example, a user might need to navigate to a particular location in the world, manipulate some aspect of the world at that location, then move to another location, etc. More work needs to be done on transitioning between these tasks.

The group also felt that Avatar Control should be added to the list of common tasks, given the prevalence of low-cost technical methods for capturing user posture, and the need to convey that information to other immersed users. Another possible task that was viewed as becoming more important going forward, particularly in the mobile domain, was Information Search. Given how much users today rely on accessing information from the Internet, the varied sources for this information, and the wide range of types of data (e.g., images, video, social network feeds, etc.), finding ways of searching efficiently in immersive worlds is an open, interesting, and relatively unexplored problem.

In sum, this was a very productive breakout session. It became clear that research in 3DUI continues to be a hot field due to the growing number of users, usage scenarios, and low-cost equipment entering the market. Also, some future areas of focus emerged, leading to excitement for continuing to move the field of 3DUI forward.

References

1 Bowman, D.A., Kruijff, E., Laviola, Jr., J.J., Poupyrev, I., (2004) 3D User Interfaces: Theory and Practice, Addison-Wesley Professional, ISBN-13: 978-0201758672.

4.4 Avatars in Virtual Reality

Betty Mohler (MPI für biologische Kybernetik – Tübingen, DE)

Avatars are an increasingly popular research topic in the field of virtual reality. The first 20-30 minutes of our discussion was spent discussing exactly what participants meant by "Avatars" and it was clear that our definitions and needs for virtual humans fell into several categories. Avatars are often defined as digital models of people that either look or behave like the users they represent (see [1]). However, other terms like virtual humans (virtual characters that try to represent a human as close in fidelity as possible) or social agents (virtual characters that fulfill a certain purpose through artificial intelligence) are also often referred to as avatars. Avatars can be achieved in multiple ways, i.e. video based capture[3], pre-made avatars experienced as the user's own body due to first-person perspective and visual-motor or visual-tactile stimulation (i.e. in [2]) and physical projections of video captured data[4]. These are just a few of the many manifestations of avatars in virtual reality.

In order to achieve high-fidelity virtual agents that act in a human way many problems need to be solved by a multi-disciplinary research group. Virtual social agents must be

able to move like humans, have casual conversation, appear intelligent, be interactive, be both reactive and proactive (specific to the user), be empathetic, perform certain functions, follow basic rules of proxemics, receive and give sensory feedback (visual, tactile, auditory). Some of the most promising applications for avatars and social agents in virtual reality are: telepresence, ergonomics/simulation, training, teaching and education, medical and health, basic science (understanding human behavior, see [5]), and of course gaming and entertainment.

In this discussion time we had three breakout groups where we tried to define grand challenge examples for avatar research. One group discussed the challenges involved with the ability to remotely care for an elderly parent or remotely put your child to bed (as a second parent). These challenges involve communicating face to face, physical interaction (to comfort/support, to help with household tasks), observe monitor mental and physical health signs, the believable presence of the remote parent to a child (visual, voice, size) and the ability to embody a remote avatar. Another group discussed a scenario for avatars in the medical health profession and education of medical professionals, specifically where limited discourse is occurring. Important to these scenarios are the ability to build trust, convey empathy and have confidence in the sometimes uncertain or emotional information that is being shared. Finally, another group considered the challenge of being able to have a portable self-representation which could be brought into the virtual reality application you are using. The challenges here are system challenges of having a standard for virtual reality with regard to model, animation method and ethical issues with regard to data security. Specifically this group considered how the data for individual avatars might be collected, e.g. cameras only, motion capture suits, physiological measures such as heart rate, skin conductance and brain waves. Specifically, the question was raised: Which measures help increase fidelity and which ones go ethically too far?

References

1 Maimone, A, X Yang, N Dierk, A State, M Dou, H Fuchs, General-Purpose Telepresence with Head-Worn Optical See-Through Displays and Projector-Based Lighting Control, IEEE VR 2013, Orlando, Florida, March 16–23, 2013.

2 Mel Slater, Bernhard Spanlang, Maria V Sanchez-Vives, Olaf Blanke (2010) First person experience of body transfer in virtual reality. PLoS ONE 5: 5. 05

3 Beck, S., Kunert, A., Kulik, A., Froehlich B. Immersive Group-to-Group Telepresence (Best Paper Award) IEEE Transactions on Visualization and Computer Graphics, 19(4):616-25, March 2013 (Proceedings of IEEE Virtual Reality 2013, Orlando, Florida).

4 Peter Lincoln, Greg Welch, Andrew Nashel, Andrei State, Adrian Ilie, and Henry Fuchs. Animatronic Shader Lamps Avatars. Virtual Reality (Springer), special issue on Augmented Reality, pp. 1–14, 2010. (see also: http://www.cs.unc.edu/~welch/media/pdf/Lincoln2009ac.pdf).

5 Mohler BJ, Creem-Regehr SH, Thompson WB and Bülthoff HH (June-2010) The Effect of Viewing a Self-Avatar on Distance Judgments in an HMD-Based Virtual Environment Presence: Teleoperators and Virtual Environments 19(3) 230-242.

6 Ishiguro, Hiroshi, Humanoid Robot: http://www.youtube.com/watch?v=uD1CdjlrTBM

4.5 Scientific Visualization and VR

Torsten Kuhlen (RWTH Aachen, DE)

Since its hype in the early 90's, Virtual Reality has undoubtedly been adopted as a useful tool in a variety of application domains, e.g. product development, training, and psychology. Yet, one of the proclaimed killer applications of the early days, namely the use of VR technology as an interface for advanced data analysis and visualization solutions, has yet to realize its full potential. While there have been some successes in the research arena, the wide-spread adoption of immersive visualization by domain scientists still hasn't come about. Starting with Andries van Dam's seminal call to action, already published in 2000 in the Computer Graphics & Applications Journal, 14 Dagstuhl Seminar participants discussed several key requirements and challenges for the future development of VR-based visualization tools, specifically targeting aspects of performance, utility, and usability. All in all, the discussion group identified two kinds of primary deficits – interface and system shortfalls. It turns out that VR interfaces developed so far are not well tailored to specific problems and tasks. Actually, domain experts from different areas think differently and have different ideas what they want to do with their data.

Existing VR interfaces do not reflect this issue sufficiently. In terms of system deficits, the group realizes that the available VR tools are not flexible or adaptable enough. Also, they lack interoperability and interactivity/performance. Most of the available frameworks concentrate on real-time rendering only, while for a fully explorative data analysis, an interactivity of the whole visualization pipeline is mandatory, including feature extraction and mapping. Here, we run into a fundamental dilemma for immersive visualization: while its biggest potential arguably is its interactivity, the amount of data that has to be analyzed in any realistically complex application oftentimes exceeds the limits of interactive processing.To make VR more visible and accepted in the communities, success stories should be gathered as a first step, to demonstrate that there already exists a considerable amount of examples where VR tools could in fact assist domain scientists in understanding their data. In the midterm, the VR community should strive to develop better interfaces, tailored to the domain scientists' needs. Thereby, interface development should not only concentrate on large VR systems like CAVEs, PowerWalls etc. Instead, by using commodity I/O devices, VR solutions should be brought to the scientists' offices. In the longer term, a large SW project is most probably necessary to develop powerful VRVis tools. Since VR comes along with quite specific requirements in terms of interaction, latency etc., it does not seem appropriate to just extend existing traditional visualization frameworks. To solve the interactivity-complexity dilemma, a suitable VR-based SW framework will possibly have to leverage supercomputing resources in an interactive way.

4.6 Characterising Interactions in Virtual (and/or Real) Environments

Paul Milgram (University of Toronto, CA)

Our discussion group was motivated by the contention that many researchers in our (extended) community are working on similar or related problems without realising it, while many researchers in our (extended) community think that they are working on similar or related problems that are actually rather different. A number of approaches were discussed about the challenge of how one can go about effectively modelling the various ways in which humans can interact with elements of virtual environments, as well as (in recognition of the importance of mixed reality) with real environments.

A distinction was made between top-down and bottom-up approaches. The former, which are characterised most effectively by parsimonious simplicity, can serve as a powerful means of revealing high level interactions, primarily by means of, among other things, qualitative representations, conceptual relationships, and metaphors. The latter bottom-up models are typically more detailed and serve as invaluable practical tools for describing and developing system architecture.

The conclusions reached included: (a) a consensus that it is indeed a useful exercise to attempt to define a framework for characterising such interactions; (b) no one framework is likely to be adequate for characterising all possible interactions; (c) top-down / conceptual / metaphorical frameworks are typically centred explicitly on users and thus can be difficult / inadequate / impractical for characterising all aspects and components of primarily complex multidimensional systems; (d) bottom-up / algorithmic / structural models are indispensable for software design and development, but are not necessarily useful for elucidating global similarities and differences among diverse interaction systems.

4.7 Unconventional Mixed Environments

Freiherr von Lukas, Uwe; Quarles, John; Staadt, Oliver

The primary objective of the Unconventional Mixed Environments (UCME) session was to discuss the potential for augmented, mixed, or virtual environments whose surrounding real environment is unconventional, e.g., underwater, zero gravity, or extreme pressure. The applications of UCME are potentially very impactful, including rehabilitation, astronaut training, and diver navigation, respectively. However, most UCME applications have been unrealized, perhaps due to limitations in technology and gaps in science.

Currently, there are only a few specific examples of UCMEs, including underwater AR Blum et.al, Zero-G experiments for AR interaction Markov-Vetter et al., and VR installations with additional wind or heat sources. Further installations have probably been done in the military area but have not been published.

In the discussion, we expanded the definition of UCME to two different types:

1. traditional VR or AR installations that shall be situated in a untypical environment
2. VR or AR installations with additional modalities to produce an illusion beyond a classical VR/AR experience (i.e. video, audio, force feedback)

In both cases, we are looking on physical effects such as pressure, temperature, acceleration, density of the surrounding media or flows of media (e.g. wind). Such UCME allow for a variety of specific applications, including training or assistance for divers or astronauts, physical rehabilitation or pain therapy as well as environments for experimenting human behavior (or animal behavior) under special conditions. The implementation of UCME brings up various technical challenges. First of all we have to find robust solutions to make the mixed reality devices resistant to the surrounding environment (water, pressure, extreme temperature, ...). Most of the display systems and sensors used in mixed environments today are designed to work in air. Changing the surrounding medium to liquids, we at least have to adapt the systems or even find completely new approaches. Depending on the state of the art, some of them can be solved with existing technology. However, some of the technical problems will lead to research in mechanical engineering, material science etc. In any case, UCME will have to be addressed by multidisciplinary teams that bring in knowledge from marine technology or aeronautics. For example, it is unknown how to track users reliably underwater.

Besides the technical challenges there are several interesting scientific challenges regarding human aspects. The previous research on how humans perceive and interact with augmented, mixed, and virtual realities (mainly based on the visual channel) must be extended. After understanding the influence of other surrounding media or different modalities for perception of virtual worlds, this knowledge can be applied to build the next generation of more sophisticated UCMEs.

References

1 Blum, Lisa, Wolfgang Broll, and Stefan Müller. "Augmented reality under water." SIGGRAPH'09: Posters. ACM, 2009.

2 Markov-Vetter, D., Moll, E., Staadt, O. "Verifying Sensorimotoric Coordination of Augmented Reality Selection under Hyper-and Microgravity" International Journal of Advanced Computer Science 3 (5) 2013

Participants

- Carlos Andujar
UPC – BarcelonaTech, ES
- Steffi Beckhaus
Eppstein, DE
- Roland Blach
FhG IAO – Stuttgart, DE
- Wolfgang Broll
TU Ilmenau, DE
- Pere Brunet
UPC – BarcelonaTech, ES
- Guido Brunnett
TU Chemnitz, DE
- Sabine Coquillart
INRIA Grenoble
Rhône-Alpes, FR
- Carolina Cruz-Neira
University of Louisiana at
Lafayette, US
- Ralf Dörner
Hochschule RheinMain –
Wiesbaden, DE
- Steven K. Feiner
Columbia University, US
- Uwe Freiherr von Lukas
FhG IGD – Rostock, DE
- Bernd Fröhlich
Bauhaus-Universität Weimar, DE
- Henry Fuchs
University of North Carolina at
Chapel Hill, US
- Martin Göbel
Hochschule Bonn-Rhein-Sieg, DE
- Raphael Grasset
TU Graz, AT
- Jens Herder
FH Düsseldorf, DE
- Tobias Höllerer
University of California –
Santa Barbara, US
- Charles E. Hughes
University of Central Florida –
Orlando, US
- Masahiko Inami
Kaio University, JP
- Victoria Interrante
University of Minnesota –
Duluth, US
- Bernhard Jung
TU Bergakademie Freiberg, DE
- Panagiotis D. Kaklis
National Technical University of
Athens, GR
- Marcelo Kallmann
Univ. of California – Merced, US
- Yoshifumi Kitamura
Tohoku University, JP
- Kiyoshi Kiyokawa
Osaka University, JP
- Gudrun Klinker
TU München, DE
- Ernst Kruijff
Hochschule Bonn-Rhein-Sieg, DE
- Torsten Kuhlen
RWTH Aachen, DE
- Marc Erich Latoschik
Universität Würzburg, DE
- Anatole Lecuyer
INRIA Rennes-Bretagne
Atlantique, FR
- Robert W. Lindeman
Worcester Polytechnic Inst., US
- Paul Milgram
University of Toronto, CA
- Mark Mine
Walt Disney Imagineering, US
- Betty Mohler
MPI für biologische Kybernetik –
Tübingen, DE
- Tabitha C. Peck
Duke University, US
- Jerome Perret
Haption – Aachen, DE
- John Quarles
University of Texas at San
Antonio, US
- Christian Sandor
Univ. of South Australia, AU
- Dieter Schmalstieg
TU Graz, AT
- Andreas Simon
FH Nordwestschweiz, CH
- Oliver Staadt
Universität Rostock, DE
- Anthony Steed
University College London, GB
- Jeanine Stefanucci
University of Utah, US
- Frank Steinicke
Universität Würzburg, DE
- Susumu Tachi
Kaio University, JP
- Robert van Liere
CWI – Amsterdam, NL
- Gregory F. Welch
University of Central Florida –
Orlando, US
- Gabriel Zachmann
Universität Bremen, DE

Report from Dagstuhl Seminar 13251

Parallel Data Analysis

Edited by

Artur Andrzejak[1], Joachim Giesen[2], Raghu Ramakrishnan[3], and Ion Stoica[4]

1 **Universität Heidelberg, DE,** artur@uni-hd.de
2 **Universität Jena, DE,** joachim.giesen@uni-jena.de
3 **Microsoft Cloud Information Services Laboratory – Redmond, US,** raghu@microsoft.com
4 **University of California – Berkeley, US,** istoica@cs.berkeley.edu

Abstract

This report documents the program and the outcomes of Dagstuhl Seminar 13251 "Parallel Data Analysis" which was held in Schloss Dagstuhl – Leibniz Center for Informatics from June 16th 2013 to June 21st 2013. During the seminar, participants presented their current research and ongoing work, and open problems were discussed. The first part of this document describes seminar goals and topics, while the remainder gives an overview of the contents discussed during this event. Abstracts of a subset of the presentations given during the seminar are put together in this paper. Links to extended abstracts or full papers are provided, if available.

Seminar 16.–21. June, 2013 – www.dagstuhl.de/13251
1998 ACM Subject Classification H.2.8 Database Applications (data mining), I.2.6 Learning, I.5 Pattern Recognition, C.2.4 Distributed Systems (Distributed applications), C.4 Performance of systems
Keywords and phrases data analysis, machine learning, parallel processing, distributed computing, software frameworks
Digital Object Identifier 10.4230/DagRep.3.6.67

1 Executive Summary

Artur Andrzejak
Joachim Giesen
Raghu Ramakrishnan
Ion Stoica

Motivation and goals

Parallel data analysis accelerates the investigation of data sets of all sizes, and is indispensable when processing huge volumes of data. The current ubiquity of parallel hardware such as multi-core processors, modern GPUs, and computing clusters has created an excellent environment for this approach. However, exploiting these computing resources effectively requires significant efforts due to the lack of mature frameworks, software, and even algorithms designed for data analysis in such computing environments.

As a result, parallel data analysis is often being used only as the last resort, i.e., when the data size becomes too big for sequential data analysis, and it is hardly ever used for

analyzing small and medium-sized data sets though it could be also beneficial for there, i.e., by cutting compute time down from hours to minutes or even making the data analysis process interactive. The barrier of adoption is even higher for specialists from other areas such as sciences, business, and commerce. These users often have to make do with slower, yet much easier to use sequential programming environments and tools, regardless of the data size.

The seminar participants have tried to address these challenges by focusing on the following goals:

- Providing user-friendly parallel programming paradigms and cross-platform frameworks or libraries for easy implementation and experimentation.
- Designing efficient and scalable parallel algorithms for machine learning and statistical analysis in connection with an analysis of use cases.

The program

The seminar program consisted of individual presentations on new results and ongoing work, a plenary session, as well as work in two working groups. The primary role of the focus groups was to foster the collaboration of the participants, allowing cross-disciplinary knowledge sharing and insights. Work in one group is still ongoing and targets as a result a publication in a magazine.

The topics of the plenary session and the working groups were the following ones:

- Panel “From Big Data to Big Money’
- Working group “A”: Algorithms and applications
- Working group “P”: Programming paradigms, frameworks and software.

2 Table of Contents

3 Abstracts of Selected Talks

3.1 Incremental-parallel Learning with Asynchronous MapReduce

Artur Andrzejak (Universität Heidelberg, DE)

Joint work of Artur Andrzejak, Joos-Hendrik Böse, Joao Bartolo Gomes, Mikael Högqvist
Main reference J.-H. Böse, A. Andrzejak, M. Högqvist, "Beyond Online Aggregation: Parallel and Incremental Data Mining with Online MapReduce," in Proc. of the 2010 Workshop on Massive Data Analytics on the Cloud (MDAC'10), 6 pp., ACM, 2010.
URL http://dx.doi.org/10.1145/1779599.1779602

MapReduce paradigm for parallel processing has turned suitable for implementing a variety of algorithms within the domain of machine learning. However, the original design of this paradigm suffers under inefficiency in case of iterative computations (due to repeated data reads from I/O) and inability to process streams or output preliminary results (due to a barrier / sync operation between map and reduce).

In the first part of this talk we propose a framework which modifies the MapReduce paradigm in twofold ways [1]. The first modification removes the barrier / sync operation, allowing reducers to process (and output) preliminary or streaming data. The second change is the mechanism to send any messages from reducers "back" to mappers. The latter property allows efficient iterative processing, as data (once read from disk or other I/O) can be kept in the main memory by map tasks, and reused in subsequent computation phases (usually, each phase being triggered by new messages/data from the reducer). We evaluate this architecture and its ability to produce preliminary results and process streams by implementing several machine learning algorithms. These include simple "one pass" algorithms like linear regression or Naive Bayes. A more advanced example is a parallel – incremental (i.e. online) version of the k-means clustering algorithm.

In the second part we focus on the issue of parallel detection of concept drift in context of classification models. We propose Online Map-Reduce Drift Detection Method (OMR-DDM) [2]. Also here our modified MapReduce framework is used. To this end, we extend the approach introduced in [3]. This is done by parallelizing training of an incremental classifier (here Naive Bayes) and the partial evaluation of its momentarily accuracy. An experimental evaluation shows that the proposed method can accurately detect concept drift while exploiting parallel processing. This paves the way to obtaining classification models which consider concept drift on massive data.

References

1 Joos-Hendrik Böse, Artur Andrzejak, Mikael Högqvist. *Beyond Online Aggregation: Parallel and Incremental Data Mining with Online MapReduce*. ACM MDAC 2010, Raleigh, NC, 2010.
2 Artur Andrzejak, Joao Bartolo Gomes. *Parallel Concept Drift Detection with Online MapReduce*. KDCloud 2012 at ICDM 2012, 10 December 2012, Brussels, Belgium.
3 João Gama and Pedro Medas and Gladys Castillo and Pedro Rodrigues. *Learning with drift detection*. Advances in Artificial Intelligence, 2004, pages 66–112, 2004

3.2 Scaling Up Machine Learning

Ron Bekkerman (Carmel Ventures – Herzeliya, IL)

Joint work of Bekkerman, Ron; Bilenko, Mikhail; Langford, John
Main reference R. Bekkerman, M. Bilenko, J. Langford, (eds.), "Scaling Up Machine Learning," Cambridge University Press, January 2012
URL http://www.cambridge.org/us/academic/subjects/computer-science/pattern-recognition-and-machine-learning/scaling-machine-learning-parallel-and-distributed-approaches

In this talk, I provide an extensive introduction to parallel and distributed machine learning. I answer the questions "How actually big is the big data?", "How much training data is enough?", "What do we do if we don't have enough training data?", "What are platform choices for parallel learning?" etc. Over an example of k-means clustering, I discuss pros and cons of machine learning in Pig, MPI, DryadLINQ, and CUDA.

3.3 Efficient Co-Processor Utilization in Database Query Processing

Sebastian Breß (Otto-von-Guericke-Universität Magdeburg, DE)

Joint work of Sebastian Breß, Felix Beier, Hannes Rauhe, Kai-Uwe Sattler, Eike Schallehn, and Gunter Saake
Main reference S. Breß, F. Beier, H. Rauhe, K.-U. Sattler, E. Schallehn, G. Saake, "Efficient Co-Processor Utilization in Database Query Processing," Information Systems, 38(8):1084–1096, 2013.
URL http://dx.doi.org/10.1016/j.is.2013.05.004

Co-processors such as GPUs provide great opportunities to speed up database operations by exploiting parallelism and relieving the CPU. However, distributing a workload on suitable (co-)processors is a challenging task, because of the heterogeneous nature of a hybrid processor/co-processor system. In this talk, we discuss current problems of database query processing on GPUs and present our decision model, which distributes a workload of operators on all available (co-)processors. Furthermore, we provide an overview of how the decision model can be used for hybrid query optimization.

References

1 S. Breß, F. Beier, H. Rauhe, K.-U. Sattler, E. Schallehn, and G. Saake. Efficient Co-Processor Utilization in Database Query Processing. *Information Systems*, 38(8):1084–1096, 2013.
2 S. Breß, I. Geist, E. Schallehn, M. Mory, and G. Saake. A Framework for Cost based Optimization of Hybrid CPU/GPU Query Plans in Database Systems. *Control and Cybernetics*, 41(4):715–742, 2012.

3.4 Analytics@McKinsey

Patrick Briest (McKinsey&Company – Düsseldorf, DE)

To successfully capture value from advanced analytics, businesses need to combine three important building blocks: Creative integration of internal and external data sources and

the ability to filter relevant information lays the foundation. Predictive and optimization models striking the right balance between complexity and ease of use provide the means to turn data into insights. Finally, a solid embedding into the organizational processes via simple, useable tools turns insights into impactful frontline actions.

This talk gives an overview of McKinsey's general approach to big data and advanced analytics and presents several concrete examples of how advanced analytics are applied in practice to business problems from various different industries.

3.5 A Data System for Feature Engineering

Michael J. Cafarella (University of Michigan – Ann Arbor, US)

Joint work of Anderson, Michael; Antenucci, Dolan; Bittorf, Victor; Burgess, Matthew; Cafarella, Michael J.; Kumar, Arun; Niu, Feng; Park, Yongjoo; Ré, Christopher; Zhang, Ce
Main reference M. Anderson, D. Antenucci, V. Bittorf, M. Burgess, M.J. Cafarella, A. Kumar, F. Niu, Y. Park, C. Ré, C. Zhang, "Brainwash: A Data System for Feature Engineering," in Proc. of the 6th Biennial Conf. on Innovative Data Systems Research (CIDR'13), 4 pp. , 2013.
URL http://www.cidrdb.org/cidr2013/Papers/CIDR13_Paper82.pdf

Trained systems, such as Web search, recommendation systems, and IBM's Watson question answering system, are some of the most compelling in all of computing. However, they are also extremely difficult to construct. In addition to large datasets and machine learning, these systems rely on a large number of machine learning features. Engineering these features is currently a burdensome and time-consuming process.

We introduce a datasystem that attempts to ease the task of feature engineering. By assuming that even partially-written features are successful for some inputs, we can attempt to execute and benefit from user code that is substantially incorrect. The system's task is to rapidly locate relevant inputs for the user- written feature code with only implicit guidance from the learning task. The resulting system enables users to build features more rapidly than would otherwise be possible.

3.6 Extreme Data Mining: Global Knowledge without Global Communication

Giuseppe Di Fatta (University of Reading, GB)

Joint work of Di Fatta, Giuseppe; Blasa, Francesco; Cafiero, Simone; Fortino, Giancarlo
Main reference G. Di Fatta, F. Blasa, S. Cafiero, G. Fortino. "Fault tolerant decentralised k-Means clustering for asynchronous large-scale networks," Journal of Parallel and Distributed Computing, Vol. 73, Issue 3, March 2013, pp. 317–329, 2013.
URL http://dx.doi.org/10.1016/j.jpdc.2012.09.009

Parallel Data Mining in very large and extreme-scale systems is hindered by the lack of scalable and fault tolerant global communication and synchronisation methods. Epidemic protocols are a type of randomised protocols which provide statistical guarantees of accuracy and consistency of global aggregates in decentralised and asynchronous networks. Epidemic K-Means is the first data mining protocol which is suitable for very large and extreme-scale systems, such as Peer-to-Peer overlay networks, the Internet of Things and exascale

supercomputers. This distributed and fully-decentralised K-Means formulation provides a clustering solution which can approximate the solution of an ideal centralised algorithm over the aggregated data as closely as desired. A comparative performance analysis with the state of the art sampling methods is presented.

3.7 Parallelization of Machine Learning Tasks by Problem Decomposition

Johannes Fürnkranz (TU Darmstadt, DE)

Joint work of Fürnkranz, Johannes; Hüllermeier, Eyke

In this short presentation I put forward the idea that parallelization can be achieved by decomposing a complex machine learning problem into a series of simpler problems than can be solved independently, and collectively provide the answer to the original problem. I illustrate this on the task of pairwise classification, which solves a multi-class classification problem by reducing it to a set of binary classification problems, one for each pair of classes. Similar decompositions can be applied to problems like preference learning, ranking, multilabel classification, or ordered classification. The key advantage of this approach is that it gives many small problems, the main disadvantage is that the number of examples that have to be distributed over multiple cores increases n-fold.

3.8 Sclow Plots: Visualizing Empty Space

Joachim Giesen (Universität Jena, DE)

Joint work of Giesen, Joachim; Kühne, Lars; Lucas, Philipp

Scatter plots are mostly used for correlation analysis, but are also a useful tool for understanding the distribution of high-dimensional point cloud data. An important characteristic of such distributions are clusters, and scatter plots have been used successfully to identify clusters in data. Another characteristic of point cloud data that has received less attention are regions that contain no or only very few data points. We show that augmenting scatter plots by projections of flow lines along the gradient vector field of the distance function to the point cloud reveals such empty regions or voids. The augmented scatter plots, that we call sclow plots, enable a much better understanding of the geometry underlying the point cloud than traditional scatter plots.

3.9 Financial and Data Analytics with Python

Yves J. Hilpisch (Visixion GmbH, DE)

Main reference Y.J. Hilpisch, "Derivatives Analytics with Python – Data Analysis, Models, Simulation, Calibration, Hedging," Visixion GmbH.
URL http://www.visixion.com/?page_id=895

The talk illustrates, by the means of concrete examples, how Python can help in implementing efficient, interactive data analytics. There are a number of libraries available, like pandas or PyTables, that allow high performance analytics of e.g. time series data or out-of-memory data. Examples shown include financial time series analytics and visualization, high frequency data aggregation and analysis and parallel calculation of option prices via Monte Carlo simulation. The talk also compares out-of-memory analytics using PyTables with in-memory analytics using pandas.

Continuum Analytics specializes in Python-based Data Exploration & Visualization. It is engaged in a number of Open Source projects like Numba (just-in-time compiling of Python code) or Blaze (next-generation disk-based, distributed arrays for Python). It also provides the free Python distribution Anaconda for scientific and enterprise data analytics.

3.10 Convex Optimization for Machine Learning Made Fast and Easy

Soeren Laue (Universität Jena, DE)

Joint work of Giesen, Joachim; Mueller, Jens; Laue, Soeren

In machine learning, solving convex optimization problems often poses an efficiency vs. convenience trade-off. Popular modeling languages in combination with a generic solver allow to formulate and solve these problems with ease, however, this approach does typically not scale well to larger problem instances. In contrast to the generic approach, highly efficient solvers consider specific aspects of a concrete problem and use optimized parameter settings. We describe a novel approach that aims at achieving both goals at the same time, namely, the ease of use of the modeling language/generic solver combination, while generating production quality code that compares well with specialized, problem specific implementations. We call our approach a generative solver for convex optimization problems from machine learning (GSML). It outperforms state-of-the-art approaches of combining a modeling language with a generic solver by a few orders of magnitude.

3.11 Interactive, Incremental, and Iterative Dataflow with Naiad

Frank McSherry (Microsoft – Mountain View, US)

Joint work of McSherry, Frank; Murray, Derek; Isaacs, Rebecca; Isard, Michael
URL http://research.microsoft.com/naiad/

This talk will cover a new computational frameworks supported by Naiad, differential dataflow, that generalizes standard incremental dataflow for far greater re-use of previous results when collections change. Informally, differential dataflow distinguishes between the multiple reasons a collection might change, including both loop feedback and new input data, allowing a system to re-use the most appropriate results from previously performed work when an incremental update arrives. Our implementation of differential dataflow efficiently executes queries with multiple (possibly nested) loops, while simultaneously responding with low latency to incremental changes to the inputs. We show how differential dataflow enables orders of magnitude speedups for a variety of workloads on real data, and enables new analyses previously not possible in an interactive setting.

3.12 Large Scale Data Analytics: Challenges, and the role of Stratified Data Placement

Srinivasan Parthasarathy (Ohio State University, US)

Joint work of Parthasarathy, Srinivasan; Wang, Ye; Chakrabarty, Aniket; Sadayappan, P
Main reference Y. Wang, S. Parthasarathy, P. Sadayappan, "Stratification driven placement of complex data: A framework for distributed data analytics," in Proc. of IEEE 29th Int'l Conf. on Data Engineering (ICDE'13), pp. 709–720, IEEE, 2013.
URL http://dx.doi.org/10.1109/ICDE.2013.6544868

With the increasing popularity of XML data stores, social networks and Web 2.0 and 3.0 applications, complex data formats, such as trees and graphs, are becoming ubiquitous. Managing and processing such large and complex data stores, on modern computational eco-systems, to realize actionable information efficiently, is daunting. In this talk I will begin with discussing some of these challenges. Subsequently I will discuss a critical element at the heart of this challenge relates to the placement, storage and access of such tera- and peta-scale data. In this work we develop a novel distributed framework to ease the burden on the programmer and propose an agile and intelligent placement service layer as a flexible yet unified means to address this challenge. Central to our framework is the notion of stratification which seeks to initially group structurally (or semantically) similar entities into strata. Subsequently strata are partitioned within this eco-system according to the needs of the application to maximize locality, balance load, or minimize data skew. Results on several real-world applications validate the efficacy and efficiency of our approach.

3.13 Big Data @ Microsoft

Raghu Ramakrishnan (Microsoft CISL, Redmond, WA, US)

Joint work of Raghu Ramakrishnan; CISL team at Microsoft

The amount of data being collected is growing at a staggering pace. The default is to capture and store any and all data, in anticipation of potential future strategic value, and vast amounts of data are being generated by instrumenting key customer and systems touchpoints. Until recently, data was gathered for well-defined objectives such as auditing, forensics, reporting and line-of-business operations; now, exploratory and predictive analysis is becoming ubiquitous. These differences in data scale and usage are leading to a new generation of data management and analytic systems, where the emphasis is on supporting a wide range of data to be stored uniformly and analyzed seamlessly using whatever techniques are most appropriate, including traditional tools like SQL and BI and newer tools for graph analytics and machine learning. These new systems use scale-out architectures for both data storage and computation.

Hadoop has become a key building block in the new generation of scale-out systems. Early versions of analytic tools over Hadoop, such as Hive and Pig for SQL-like queries, were implemented by translation into Map-Reduce computations. This approach has inherent limitations, and the emergence of resource managers such as YARN and Mesos has opened the door for newer analytic tools to bypass the Map-Reduce layer. This trend is especially significant for iterative computations such as graph analytics and machine learning, for which Map-Reduce is widely recognized to be a poor fit. In this talk, I will examine this architectural trend, and argue that resource managers are a first step in re-factoring the early implementations of Map-Reduce, and that more work is needed if we wish to support a variety of analytic tools on a common scale-out computational fabric. I will then present REEF, which runs on top of resource managers like YARN and provides support for task monitoring and restart, data movement and communications, and distributed state management. Finally, I will illustrate the value of using REEF to implement iterative algorithms for graph analytics and machine learning.

3.14 Berkeley Data Analytics Stack (BDAS)

Ion Stoica (University of California – Berkeley, US)

One of the most interesting developments over the past decade is the rapid increase in data; we are now deluged by data from on-line services (PBs per day), scientific instruments (PBs per minute), gene sequencing (250GB per person) and many other sources. Researchers and practitioners collect this massive data with one goal in mind: extract "value" through sophisticated exploratory analysis, and use it as the basis to make decisions as varied as personalized treatment and ad targeting. Unfortunately, today's data analytics tools are slow in answering even simple queries, as they typically require to sift through huge amounts of data stored on disk, and are even less suitable for complex computations, such as machine learning algorithms. These limitations leave the potential of extracting value of big data unfulfilled.

To address this challenge, we are developing BDAS, an open source data analytics stack that provides interactive response times for complex computations on massive data. To achieve this goal, BDAS supports efficient, large-scale in-memory data processing, and allows users and applications to trade between query accuracy, time, and cost. In this talk, I'll present the architecture, challenges, early results, and our experience with developing BDAS. Some BDAS components have already been released: Mesos, a platform for cluster resource management has been deployed by Twitter on +6,000 servers, while Spark, an in-memory cluster computing frameworks, is already being used by tens of companies and research institutions.

3.15 Scalable Data Analysis on Clouds

Domenico Talia (University of Calabria, IT)

URL http://gridlab.dimes.unical.it

This talk presented a Cloud-based framework designed to program and execute parallel and distributed data mining applications: The Cloud Data Mining Framework. It can be used to implement parameter sweeping applications and workflow-based applications that can be programmed through a graphical interface and trough a script-based interface that allow to compose a concurrent data mining program to be run on a Cloud platform. We presented the main system features and its architecture. In the Cloud Data Mining framework each node of a workflow is a service, so the application is composed o a collection of Cloud services.

3.16 Parallel Generic Pattern Mining

Alexandre Termier (University of Grenoble, FR)

Joint work of Termier, Alexandre; Negrevergne, Benjamin; Mehaut, Jean-Francois; Rousset, Marie-Christine
Main reference B. Negrevergne, A. Termier, M.-C. Rousset, J.-F. Méhaut, "ParaMiner: a generic pattern mining algorithm for multi-core architectures," iData Mining and Knowledge Discovery, April 2013, Springer, 2013.
URL http://dx.doi.org/10.1007/s10618-013-0313-2

Pattern mining is the field of data mining concerned with finding repeating patterns in data. Due to the combinatorial nature of the computations performed, it requires a lot of computation time and is therefore an important target for parallelization. In this work we show our parallelization of a generic pattern mining algorithm, and how the pattern definition influes on the parallel scalability. We also show that the main limiting factor is in most cases the memory bandwidth, and how we could overcome this limitation.

3.17 REEF: The Retainable Evaluator Execution Framework

Markus Weimer (Microsoft CISL, Redmond, WA, US)

Joint work of Chun, Byung-Gon; Condie, Tyson; Curino, Carlo; Douglas, Chris; Narayanamurthy, Shravan; Ramakrishnan, Raghu; Rao, Sriram; Rosen, Joshua; Sears, Russel; Weimer, Markus

The Map-Reduce framework enabled scale-out for a large class of parallel computations and became a foundational part of the infrastructure at Web companies. However, it is recognized that implementing other frameworks such as SQL and Machine Learning by translating them into Map-Reduce programs leads to poor performance.

This has led to a refactoring the Map-Reduce implementation and the introduction of domain-specific data processing frameworks to allow for direct use of lower-level components. Resource management has emerged as a critical layer in this new scale-out data processing stack. Resource managers assume the responsibility of multiplexing fine-grained compute tasks on a cluster of shared-nothing machines. They operate behind an interface for leasing *containers*—a slice of a machine's resources (e.g., CPU/GPU, memory, disk)—to computations in an elastic fashion.

In this talk, we describe the Retainable Evaluator Execution Framework (REEF). It makes it easy to retain state in a container and reuse containers across different tasks. Examples include pipelining data between different operators in a relational pipeline; retaining state across iterations in iterative or recursive distributed programs; and passing state across different types of computations, for instance, passing the result of a Map-Reduce computation to a Machine Learning computation.

REEF supports this style of distributed programming by making it easier to: (1) interface with resource managers to obtain containers, (2) instantiate a runtime (e.g., for executing Map-Reduce or SQL) on allocated containers, and (3) establish a control plane that embodies the application logic of how to coordinate the different tasks that comprise a job, including how to handle failures and preemption. REEF also provides data management and communication services that assist with task execution. To our knowledge, this is the first approach that allows such reuse of dynamically leased containers, and offers potential for order-of-magnitude performance improvements by eliminating the need to persist state (e.g., in a file or shared cache) across computational stages.

4 Group Composition and Schedule

4.1 Participants

The seminar has brought together academic researchers and industry practitioners to foster cross-disciplinary interactions on parallel analysis of scientific and business data. The following three communities were particularly strongly represented:

- researchers and practitioners in the area of frameworks and languages for data analysis
- researchers focusing on machine learning and data mining
- practitioners analysing data of various sizes in the domains of finance, consulting, engineering, and others.

In summary, the seminar gathered 36 researchers from the following 10 countries:

Country	Number of participants
Canada	1
France	1
Germany	13
Israel	1
Italy	1
Korea	1
Portugal	1
Singapore	1
UK	1
USA	15

Most participants came from universities or state-owned research centers. However, a considerable fraction of them were affiliated with industry or industrial research centers – altogether, 13 participants. Here is a detailed statistic of the affiliations:

Industry	Institution	Country	# Participants
	Argonne National Laboratory	USA	1
	Brown University – Providence	USA	1
Yes	Carmel Ventures – Herzeliya	Israel	1
	Freie Universität Berlin	Germany	1
Yes	Institute for Infocomm Research (I2R)	Singapore	1
Yes	McKinsey & Company	Germany	1
Yes	Microsoft and Microsoft Research	USA	6
	Ohio State University	USA	1
	Otto-von-Guericke-Universität Magdeburg	Germany	1
Yes	SAP AG	Germany	2
Yes	SpaceCurve	USA	1
	Stony Brook University / SUNY Korea	USA / Korea	1
	TU Berlin	Germany	1
	TU Darmstadt	Germany	1
	Universidade do Porto	Portugal	1
	Universität Heidelberg	Germany	2
	Universität Jena	Germany	3
	University of Alberta	Canada	1
	University of Calabria	Italy	1
	University of California – Berkeley	USA	3
	University of Grenoble	France	1
	University of Michigan	USA	1
	University of Minnesota	USA	1
	University of Reading	UK	1
Yes	Visixion GmbH / Continuum Analytics	Germany	1

4.2 Complete list of talks

Monday, June 17th 2013

S1: Applications

Krishnaswamy, Shonali	Mobile & Ubiquitous DataStream Mining
Broß, Jürgen	Mining Customer Review Data
Will, Hans-Martin	Real-time Analysis of Space and Time

S2: Frameworks I

Peterka, Tom	Do-It-Yourself Parallel Data Analysis
Joseph, Anthony D.	Mesos
Zaharia, Matei	The Spark Stack: Making Big Data Analytics Interactive and Real-time

Tuesday, June 18th 2013

S3: Overview & Challenges I

Bekkerman, Ron	Scaling Up Machine Learning: Parallel and Distributed Approaches
Ramakrishnan, Raghu	Big Data @ Microsoft

S4: Overview & Challenges II

Briest, Patrick	Analytics @ McKinsey
Parthasarathy, Srinivasan	Scalable Analytics: Challenges and Renewed Bearing

S5: Frameworks II

Stoica, Ion	Berkeley Data Analytics Stack (BDAS)
Hilpisch, Yves	Financial and Data Analytics with Python
Cafarella, Michael J.	A Data System for Feature Engineering

Wednesday, June 19th 2013

S6: Visualisation and Interactivity

Giesen, Joachim	Visualizing empty space
McSherry, Frank	Interactive, Incremental, and Iterative Data Analysis with Naiad

S7: Various

Müller, Klaus	GPU-Acceleration for Visual Analytics Tasks
Laue, Soeren	Convex Optimization for Machine Learning made Fast and Easy
Di Fatta, Giuseppe	Extreme Data Mining: Global Knowledge without Global Communication

Thursday, June 20th 2013

S8: Frameworks III

Talia, Domenico	Scalable Data Analysis workflows on Clouds
Weimer, Markus	REEF: The Retainable Evaluator Execution Framework
Termier, Alexandre	Prospects for parallel pattern mining on multicores

S9: Efficiency

Andrzejak, Artur	Incremental-parallel learning with asynchronous MapReduce
Fürnkranz, Johannes	Parallelization of machine learning tasks via problem decomposition
Breß, Sebastian	Efficient Co-Processor Utilization in Database Query Processing

Participants

- Artur Andrzejak
Universität Heidelberg, DE
- Ron Bekkerman
Carmel Ventures – Herzeliya, IL
- Joos-Hendrik Böse
SAP AG – Berlin, DE
- Sebastian Breß
Universität Magdeburg, DE
- Patrick Briest
McKinsey&Company – Düsseldorf, DE
- Jürgen Broß
FU Berlin, DE
- Lutz Büch
Universität Heidelberg, DE
- Michael J. Cafarella
University of Michigan – Ann Arbor, US
- Surajit Chaudhuri
Microsoft Res. – Redmond, US
- Tyson Condie
Yahoo! Inc. – Burbank, US
- Giuseppe Di Fatta
University of Reading, GB
- Rodrigo Fonseca
Brown University, US
- Johannes Fürnkranz
TU Darmstadt, DE
- Joao Gama
University of Porto, PT
- Joachim Giesen
Universität Jena, DE
- Philipp Große
SAP AG – Walldorf, DE
- Max Heimel
TU Berlin, DE
- Yves J. Hilpisch
Visixion GmbH, DE
- Anthony D. Joseph
University of California – Berkeley, US
- George Karypis
University of Minnesota – Minneapolis, US
- Shonali Krishnaswamy
Infocomm Research – Singapore, SG
- Soeren Laue
Universität Jena, DE
- Frank McSherry
Microsoft – Mountain View, US
- Jens K. Müller
Universität Jena, DE
- Klaus Mueller
Stony Brook University, US
- Srinivasan Parthasarathy
Ohio State University, US
- Tom Peterka
Argonne National Laboratory, US
- Raghu Ramakrishnan
Microsoft Res. – Redmond, US
- Ion Stoica
University of California – Berkeley, US
- Domenico Talia
University of Calabria, IT
- Alexandre Termier
University of Grenoble, FR
- Markus Weimer
Microsoft Res. – Redmond, US
- Hans-Martin Will
SpaceCurve – Seattle, US
- Matei Zaharia
University of California – Berkeley, US
- Osmar Zaiane
University of Alberta, CA

Report from Dagstuhl Seminar 13252

Interoperation in Complex Information Ecosystems

Edited by

Andreas Harth[1], Craig A. Knoblock[2], Kai-Uwe Sattler[3], and Rudi Studer[4]

1 **KIT – Karlsruhe Institute of Technology, DE,** harth@kit.edu
2 **University of Southern California – Marina del Rey, US,** knoblock@isi.edu
3 **TU Ilmenau, DE,** kus@tu-ilmenau.de
4 **KIT – Karlsruhe Institute of Technology, DE,** studer@kit.edu

Abstract

This report documents the program and the outcomes of Dagstuhl Seminar 13252 "Interoperation in Complex Information Ecosystems".

Seminar 16.–19. June, 2013 – www.dagstuhl.de/13252
1998 ACM Subject Classification D.3.1 Formal Definitions and Theory, H.2.3 Languages, H.2.4 Systems, H.2.8 Database Applications, I.2.4 Knowledge Representation Formalisms and Methods
Keywords and phrases Information integration, System interoperation, Complex information ecosystems, Dataspaces, Linked Data, Semantic Web, Sensor networks, Restful design, Web architecture.
Digital Object Identifier 10.4230/DagRep.3.6.83

1 Executive Summary

Andreas Harth
Craig A. Knoblock
Kai-Uwe Sattler
Rudi Studer

Individuals, enterprises and policy makers increasingly rely on data to stay informed and make decisions. The amount of available digital data grows at a tremendous pace. At the same time, the number of systems providing and processing data increases, leading to complex information ecosystems with large amounts of data, a multitude of stakeholders, and a plethora of data sources and systems. Thus, there is an increasing need for integration of information and interoperation between systems.

Due to the ubiquitous need for integration and interoperation, many research communities have tackled the problem. Recent developments have established a pay-as-you-go integration model, where integration is seen as a process starting out with enabling only basic query functionality over data and iteratively spending targeted integration effort as the need for more complex queries arises. Such an ad-hoc model is in contrast to previous integration models which required the construction of a mediated schema and the integration of schema and data before any queries – even simple ones – could be answered. The move towards less rigid integration systems can be traced back to many communities: the database community established Dataspaces as a new abstraction for information integration; the Semantic Web

community provided ontologies and logic-based modelling in a web context; finally, the Web community established the Hypermedia principle which enables decentralized discovery and ad-hoc unidirectional interlinking in very large information systems.

Current systems for data integration focus on query-related aspects. However, to enable real interoperation, updates and invocation of functionality are required. Mobile applications, for example, require both access to information and functionality. We want to broaden the scope of research on data integration towards a vision of interoperation between systems (i.e., systems that no only exchange and integrate their data but also link functionality) and investigate how an iterative model can be established for the interoperation of systems.

The seminar has multiple objectives:

- to bring together researchers from these diverse communities to identify common themes and to exploit synergies,
- to develop the theoretical foundations and an understanding of architectures and methods,
- to develop a research agenda and road-map towards a vision of web-scale integration and interoperation.

Acknowledgement

The seminar has been supported by ONR Global, grant N62909-13-1-C161.

2 Table of Contents

3 Overview of Talks

3.1 Subspace Methods for Data Integration

Felix Bießmann (TU Berlin, DE)

Main reference F. Bießmann, F.C. Meinecke, A. Gretton, A. Rauch, G. Rainer, N.K. Logothetis, K.-R. Müller, "Temporal kernel CCA and its application in multimodal neuronal data analysis," Machine Learning Journal, 79(1–2):5–27, 2010.
URL http://dx.doi.org/10.1007/s10994-009-5153-3

A lot of machine learning research focuses on methods r integration of multiple data streams that are coupled via complex time-dependent and potentially non-linear dependencies. More recently we started applying these methods to web data, such as online social networks or online publishing media, in order to study aspects of temporal dynamics between web sites and users in online social networks. Our results suggest that statistical learning methods can be useful for data integration in complex information ecosystems.

Many data sets in complex information ecosystems are characterized by multiple views on data. An example could be online social networks such as Twitter: one view on each data point – a short text message called *tweet* – is the actual text of the message. Another view on this data point is the geographical information coupled to this message.

The generative model underlying our analysis approach assumes there is a hidden variable, called Z, that gives rise to all possible views on this data point. But we usually can only observe one of multiple views X, Y, ... on this hidden variable. In the Twitter example, the hidden variable would be all the semantic content of a message, and the views one can investigate are a textual realization and the geographical information. Both of these views carry complementary and shared information. The data integration problem is then to reconstruct the hidden variable from these multiple views. The reconstruction should ideally combine the textual and the geographical information. Combining this information can be useful for understanding geographical trends, trends in the text domain (i.e. topics) and how these two are related. For instance some topics could be connected to the geographical location of the users, while others are not.

An important complication that arises in the context of data integration is that different views on the hidden variables might not be coupled in a linear way. But most data integration algorithms rest on the assumption of linear couplings. Another important aspect is: Most modern data sets contain rapidly evolving high-dimensional time series. Especially when dealing with web data streams with high temporal resolution, the different views might be coupled via complex temporal dynamics.

To tackle these difficulties we developed a machine learning method, temporal kernel canonical correlation analysis (tkCCA) [1], for optimal data integration in the context of non-linearly and non-instantaneously coupled multimodal data streams. Given two (or more) multivariate time series $x \in \mathbb{R}^U$ and $y \in \mathbb{R}^V$ the method finds a projection ϕ_x and ϕ_y for each data view that projects the data into a *canonical subspace* in which x and y are maximally correlated

$$\underset{\phi_x,\phi_y}{\operatorname{argmax}} \operatorname{Corr}\left(\phi_x(y), \phi_y(y)\right).$$

The underlying assumption is that we can approximate the hidden variable by extracting a subspace that information in x and y that is maximally correlated. Working with the data in the *canonical subspace* can dramatically reduce computational complexity while preserving

all relevant information. In our future work we will extend this approach to different types of data from complex information ecosystems.

References

1 Felix Bießmann, Frank C Meinecke, Arthur Gretton, Alexander Rauch, Gregor Rainer, Nikos K Logothetis, and Klaus-Robert Müller. *Temporal kernel CCA and its application in multimodal neuronal data analysis.* Machine Learning Journal, 79(1-2):5–27, 2010.

3.2 Data Integration in Global Data Spaces

Christian Bizer (Universität Mannheim, DE)

The idea of Data Spaces [1] as a new abstraction for large-scale data management has been around for quite some time but the development of concrete methods for integrating data from such spaces was still hampered by the lack of good testbeds that enable researchers to employ their methods in large-scale, real-world settings. This situation has changed in the last years with the emergence of two global public data spaces: The Web of HTML-embedded Data and the Web of Linked Data.

1. **Web of HTML-embedded Data** [2, 3, 4]**:** This data space consists of all web sites that embed structured data into their HTML pages using markup formats such as RDFa, Microdata or Microformats. A recent analysis[1] of the Common Crawl, a large web corpus consisting of around 3 billion pages, showed that 12.3% of all web pages embed structured data. These pages originate from 2.29m websites (PLDs) among the 40.5m websites contained in the corpus (5.65%). The main topical areas of the data are people and organizations, blog- and CMS-related metadata, navigational metadata, product data, and review data, and event data. The data is structurally rather shallow but plentiful. For instance, there are 35,000 websites providing product descriptions and 16,000 websites containing business listings, both challenging test cases for identity resolution as well as data fusion methods.
2. **Web of Linked Data** [5]**:** A public global data space in which data is more structured and in which data providers ease data integration by providing integration hints in the form of RDF links is the Web of Linked Data. The Web of Linked Data contains data from a wide range of domains including geographic information, people, companies, online communities, films, music, e-government data, library data and scientific research data[2].

I think that both data spaces provide challenging requirements and are nice testbeds for developing data space integration methods. I'm interested in discussing at the workshop the research challenges that arise from global public data spaces including questions such as:

1. **Data Space Profiling:** How to describing and summarize global data spaces as well as individual data sources within these spaces? Which metrics are suitable? Which sampling strategies make sense?

[1] http://webdatacommons.org/
[2] http://lod-cloud.net/state/

2. **Schema Matching:** How to adjust schema matching methods to the specifics of global data spaces? How to make them scale to 1000s of data sources? How to enable them to take advantage of integration hints provided by the community in the form of RDF links?
3. **Identity Resolution:** How to determine the potentially large set of data sources that describe a specific real-world entity? How to make identity resolution methods scale to situations involving 1000s of data sources?
4. **Data Quality Assessment:** How to assess the quality of data within global public data spaces? Which quality dimensions matter? Which metrics can be used to assess these dimensions? Which quality-related meta-information is available? How to fuse data from large sets of data sources based on the assessment results?

References

1 M. Franklin, A. Halevy, and D. Maier. *From Databases to Dataspaces: A new Abstraction for Information Management*. SIGMOD Record 34, 4, pages 27—33, 2005.
2 H. Mühleisen and C. Bizer. *Web Data Commons – Extracting Structured Data from two Large Web Corpora*. Proceedings of LDOW 2012: Linked Data on the Web, CEUR Workshop Proceedings. CEUR-ws.org, 2012.
3 P. Mika. *Microformats and RDFa Deployment across the Web*. http://tripletalk.wordpress.com/2011/01/25/rdfa-deployment-across-the-web/, 2011.
4 P. Mika and T. Potter. *Metadata Statistics for a large Web Corpus*. Proceedings of LDOW 2012: Linked Data on the Web, CEUR Workshop Proceedings. CEUR-ws.org, 2012.
5 T. Heath and C. Bizer. *Linked Data – Evolving the Web into a Global Data Space*. Synthesis Lectures on the Semantic Web: Theory and Technology, Morgan Claypool, 2011.

3.3 A Case for Prototypes: Knowledge Representation on the Web

Stefan Decker (National University of Ireland – Galway, IE)

Motivation

Languages like OWL (Web Ontology Language) are providing means for knowledge representation on the Web using Description Logics [1], especially for the representation of Ontologies. They in particular provide means for the definition Classes. The notion of Classes and Instances has been evolved from Frame based representation languages and was first formalised by Pat Hayes in [4]. The same formalisation then has been used in the development of the theory of Description Logics. The notion of classes and instances has some properties which seem contrary to the notion of Knowledge Sharing on the Web and the development of reusable Web ontologies:

- During modeling of a particular domain the choice of what should be modeled as instances or a class sometimes seems arbitrary. Consider the creation of a knowledge bases about biological species. Should the species be modeled as classes or instances?
- Classes and Instances limit the way information can be shared. It enables only "vertical" information sharing by means of a central ontology, but provides no means of "horizontal" information sharing between peers, which seems to be a central purpose of a decentralised information system like the Web.

Prototypes

Early Frame Representation systems also deployed alternative approaches to classes and instances. [5] mentions "Prototypes: KRL, RLL, and JOSIE employ prototype frames to represent information about a typical instance of a class as opposed to the class itself and as opposed to actual instances of the class." To our knowledge, a theory of prototype based knowledge representation has not been developed. However, prototypes have been investigated in programming languages. Early examples include Self [6, 7] and more recently languages like ECMAScript [3] and its various variants (e.g., JavaScript). The basic process of reuse and sharing in prototype based programming languages is called Cloning, whereby a new object is constructed by copying the properties of an existing object (its prototype). In addition the cloned object can be modified. In some systems the resulting child object maintains an explicit link to its prototype, and changes in the prototype cause corresponding changes to be apparent in its clone. Those ideas can be adapted to enable "horizontal" Knowledge Representation on the Web.

Using Prototypes on the Web – an Example

We illustrate the usage of prototypes for Knowledge Representation on the Web with two examples.

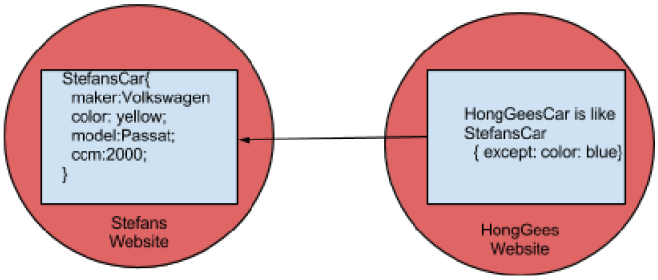

Figure 1 Horizontal information sharing on the Web.

As an example consider Figure 1: a Website (or data site) publishes a object description called "StefansCar". Another Website ("HongGees Website") is referencing StefansCar and states that HongGeesCar is like StefansCar with the change of having a different color. This enables the development of a network of horizontally shared information and the emergence of "popular" objects, which could be widely used (and therefore of "Ontologies".

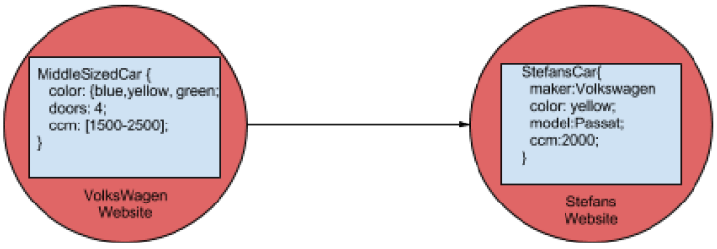

Figure 2 Emulation of Class-based Ontologies using Prototypes.

However, prototypes can also emulate the current use of class and and instances based knowledge representation languages. As Figure 2 illustrates, prototypes can be used similar

to the notion of classes and can be specialised instead of just being cloned and changed. Figure 2 shows as an examples a prototype provided on a central side (a car maker), which is being used similar to to a class by being specialised. In consequence prototypes can still provide reusable vocabularies such as FOAF or SIOC.

Prototypes – an Attempt of a Research Agenda

Using prototypes for Knowledge Representation has been fallen by the wayside and consequently to our knowledge no theory has been developed in the last years. There is a need to at least investigate the following topics:

- Formalisation of Prototypes
- Inheritance Properties
- Learning new classifications (Data Mining)
- Representation
- Large Scale Storage and Querying of Prototypes

References

1 S. Decker, D. Fensel, F. Van Harmelen, I. Horrocks, S. Melnik, M. Klein, J. Broekstra. *Knowledge Representation on the Web*. Proceedings of the 2000 Description Logic Workshop, pages 89–97, 2000.

2 C. Dony, J. Malenfant, and D. Bardou. *Classifying prototype-based programming languages*. Ivan Moore, James Noble, and Antero Taivalsaari, editors, Prototype-Based Programming, pages 17–45. Springer Verlag, 1999.

3 *Standard ECMA262 ECMAScript Language Specification*. Edition 5.1, June 2011. http://www.ecmainternational.org/publications/standards/Ecma262.htm

4 P. J. Hayes *The logic of frames*. D. Metzing, ed., Frame Conceptions and Text Understanding, Walter de Gruyter & Co., Berlin, 1979. Reprinted in Brachman & Levesque, pages 287–295, 1985.

5 P. D. Karp. *The Design Space of Frame Knowledge Representation Systems*. SRI International, 1993.

6 D. Ungar and R. B. Smith. *Self: The power of simplicity*, ACM Sigplan Notices, vol. 22. no. 12, pages 222–242, 1987.

7 R. B. Smith and D. Ungar. *Programming as an experience: The inspiration for Self*. Proceedings of ECOOP'95 – ObjectOriented Programming, 9th European Conference, Aarhus, Denmark. Lecture Notes in Computer Science Volume 952, pp 303–330, 1995.

3.4 Navigational Languages for the Web of Linked Data

Valeria Fionda (Free University of Bozen-Bolzano, IT)

Joint work of Consens, Mariano; Fionda, Valeria; Gutierrez, Claudio; Pirrò, Giuseppe

The increasing availability of structured data on the Web stimulated a renewed interest in its graph nature. The classical Web of interlinked documents is transforming into a Web of interlinked Data. In particular, the Linked Open Data project sets some informal principles for the publishing and interlinking of open data on the Web by using well-established technologies (e.g., RDF and URIs). The Web of linked data can be seen as a semantic graph where nodes represent resources and edges are labeled by RDF predicates.

Research in designing languages for accessing data in graphs and the intrinsic semantic nature of the Web of linked data suggest that graph navigation is a useful tool to address portion of this huge semantic graph. Navigation is the process of going form one point (node) in the graph to another by traversing edges. With semantic graphs, the traversal can go beyond the classical crawling that consists in traversing all the edges toward other nodes. It is possible to drive the traversal by some high-level semantic specification that encodes a reachability test, that is, the checking if from a given seed node there exists a path (defined by considering the semantics of edges) toward other nodes.

However, traditional techniques are not suitable for the Web of linked data graph. In fact the whole graph is not known a priori (as assumed by existing graph languages) but its structure has to be discovered on the fly. Hence, the notion of graph navigation has to be rethought to be useful in this new setting.

Some navigational languages have been proposed to work on *discoverable* graphs such as the Web of linked data (e.g. NautiLOD [2] and GenTLE [1]). However, most of the current navigational languages enable to specify *relevant* resources on the Web of linked data (i.e., sets of nodes in the Web of linked data) connected by a sequence of edges that match an expression but they do not provide information about the structure of the fragment where these nodes have been found. Such piece of information is crucial in some contexts such for instance citation or social networks. Hence, there is the need to augment current navigational languages with capabilities to extract *fragments* (i.e., subgraphs) of the graph being navigated besides of sets of nodes [1, 3].

References

1 M. Consens, V. Fionda, G. Pirró. GenTLE: Traversing Discoverable Graphs.Short paper at the 7th Alberto Mendelzon International Workshop on Foundations of Data Management (AMW) , 2013.
2 V. Fionda, C. Gutierrez, G. Pirró. Semantic Navigation on the Web of Data: Specification of Routes, Web Fragments and Actions. Proc. of the 21st World Wide Web Conference (WWW), pp. 281–290, 2012.
3 V. Fionda, C. Gutierrez, G. Pirró. Extracting Relevant Subgraphs from Graph Navigation. Proc. of the 11th International Semantic Web Conference (ISWC) (Posters & Demos) , 2012.

3.5 Improving Scientific Practice for Data Integration and Analysis

Yolanda Gil (University of Southern California – Marina del Rey, US)

Working with a variety of science data to support scientific discovery [8, 1], we face several major challenges.

Reducing the Cost of Data Preparation and Integration

It is estimated that 60-80% of the effort in a science project is devoted to data preparation and integration, before any science can be done. Reducing this effort is essential to accelerate the pace of discoveries. Our research focuses on several areas:

1. Reuse of data preparation software: We are looking at embedding open software practices within science communities, so that data preparation software is freely shared. We have developed semantic workflow techniques that can reason about constraints and metadata of both data and analytic steps [6], and learning approaches to extract workflow fragments that are common across workflows [3].
2. Crowdsourcing data preparation and integration through organic data sharing: We are investigating organic data sharing as a new approach to opening scientific processes so that the purpose of the data is transparent and encourages volunteer contributions from scientists [8]. An important aspect of this framework is that contributors get credit for their work, whether it is for sharing data they collect, describing data shared by others, or normalizing or integrating datasets so they can be analyzed together.
3. Leveraging the Web of Data for scientific data curation: Scientific data curation involves including metadata descriptions about world objects that may have been described elsewhere (e.g., limnology datasets could refer to lake characteristics, paleoclimatology datasets could refer to coral reef locations and species etc). We are developing user interfaces to Linked Open Data [12], and integrating them with scientific data curation systems.

Improving Provenance Practices

Scientists keep detailed provenance of their data, but in very rudimentary ways. Most of the tools that they currently use are oblivious to provenance and other metadata, which is typically managed by hand. This makes data very hard to understand by other people and by machines, and therefore its integration and reuse continues to require manual effort. The W3C PROV standard for Web provenance [9], finalized in April 2013, provides a basis for improving provenance practices, but many challenges remain:

1. Provenance-aware software: We must design data analysis software so that it uses the metadata that is available to automate data preparation and analysis, and so that it automatically creates appropriate metadata for any outputs generated [7]. Provenance-aware software can enable intelligent assistance and automation.
2. Reconstructing provenance: Provenance is typically incomplete and often incorrect, so it is important to develop approaches to make informed hypothesis about the provenance of data [11].
3. Incorporating provenance and metadata capture in science practice: Capturing provenance should be embedded in the practice of science rather than being an aside or afterthought. We are working on improving scientific publications with additional provenance, such as capturing methods as workflows [4] and improving data citations [10].

Data Science Education

A major challenge we face is the need to educate practitioners in all aspects of data science. Data science curricula are beginning to emerge that focus mostly on statistical aspects of data analytics, scale aspects concerning distributed execution, and database management. Existing curricula omit important topics such as data citation, provenance generation, and metadata tracking. Lack of awareness of these important aspects of data science is problematic, as practitioners need to address them and unfortunately end up doing so using primitive means that will not scale in the era of big data.

References

1 Deelman, E.; Duffy, C.; Gil, Y.; Marru, S.; Pierce, M.; and Wiener, G. *EarthCube Report on a Workflows Roadmap for the Geosciences*. National Science Foundation, Arlington, VA., 2012. https://sites.google.com/site/earthcubeworkflow/earthcube-workflows-roadmap

2 D. Garijo and Y. Gil. *A New Approach for Publishing Workflows: Abstractions, Standards, and Linked Data*. Proceedings of the 6th Workshop on Workflows in Support of Large-Scale Science (WORKS-11), held in conjunction with SC-11, Seattle, WA. 2011.

3 D. Garijo, O. Corcho and Y. Gil. *Detecting common scientific workflow fragments using execution provenance*. Proceedings of the International Conference on Knowledge Capture (K-CAP), Banff, Alberta, 2013.

4 D. Garijo, S. Kinnings, L. Xie, L. Xie, Y. Zhang, P. E. Bourne, and Yolanda Gil. *Quantifying Reproducibility in Computational Biology: The Case of the Tuberculosis Drugome*. To appear, 2013.

5 Gil, Y. and H. Hirsh (Eds). *Final Report of the NSF Workshop on Discovery Informatics*. National Science Foundation project report, August, 2012. http://www.discoveryinformaticsinitiative.org/diw2012

6 Gil, Y.; Gonzalez-Calero, P. A.; Kim, J.; Moody, J.; and Ratnakar, V. *A Semantic Framework for Automatic Generation of Computational Workflows Using Distributed Data and Component Catalogs*. Journal of Experimental and Theoretical Artificial Intelligence, 23(4), 2011.

7 Gil, Y.; Szekely, P.; Villamizar, S.; Harmon, T.; Ratnakar, V.; Gupta, S.; Muslea, M.; Silva, F.; and Knoblock, C. *Mind Your Metadata: Exploiting Semantics for Configuration, Adaptation, and Provenance in Scientific Workflows*. Proceedings of the Tenth International Semantic Web Conference (ISWC), Bonn, Germany. 2011.

8 Gil, Y.; Ratnakar, V.; and Hanson, P. *Organic Data Sharing: A Novel Approach to Scientific Data Sharing*. Proceedings Second International Workshop on Linked Science: Tackling Big Data (LISC), In conjunction with the International Semantic Web Conference (ISWC), Boston, MA. 2012.

9 Gil, Y.; Miles, S.; Belhajjame, K.; Deus, H.; Garijo, D.; Klyne, G.; Missier, P.; Soiland-Reyes, S.; and Zednik, S. *A Primer for the PROV Provenance Model*. World Wide Web Consortium (W3C) Technical Report, 2013.

10 A. Goodman, C. L. Borgman, K. Cranmer, Y. Gil, P. Groth, D. W. Hogg, V. Kashyap, A. Mahabal, X. Meng, A. Pepe, A. Siemiginowska, A. Slavkovic, R. Di Stefano. *Ten Simple Rules for the Care and Feeding of Scientific Data*. Submitted for publication, May 2013. https://www.authorea.com/users/23/articles/1394/

11 Groth, P.; Gil, Y.; and Magliacane, S. *Automatic Metadata Annotation through Reconstructing Provenance*. Proceedings of the Third International Workshop on the Role of Semantic Web in Provenance Management (SWPM), Heraklion, Greece, 2012.

12 Denny Vrandecic, Varun Ratnakar, Markus Krötzsch, Yolanda Gil. *Shortipedia: Aggregating and Curating Semantic Web Data*. Journal of Web Semantics, 9(3). 2011.

3.6 Supporting the Linked Data Lifecycle

Armin Haller (Australian National University, AU)

Introduction

The continuous growth of the Linked Data Web brings us closer to the original vision of the semantic Web – as an interconnected network of machine-readable resources. One of the reasons for the growth of Linked Data has been the significant progress on developing ontologies that can be used to define data in a variety of domains. The tools of choice for creating and maintaining quality-assured ontology instances (the so-called *ABox*) are still ontology editors such as Protégé. However, creating the *ABox* in an ontology editor requires some degree of understanding of RDF(s) and OWL since the user has to define to which class an individual belongs to and what are the permissible relationships between individuals. Further, as ontology editors do not separate the schema editing from the data editing, users can, for example, inadvertently make changes to the classes and relations in the ontology (the so-called TBox) while creating data. Addressing this issue, some Web publishing tools on top of Wikis, Microblogs or Content Management systems have been developed (e.g. the work discussed in [2], [6] and [3]) that allow a user to exclusively create ontology instances. However, they are mostly developed for a specific domain (i.e. specific ontologies) and often do not strictly follow OWL semantics and consequently allow the creation of an unsatisfiable ABox. Consequently, manually created, quality-assured, crowd-sourced semantic Web datasets are still largely missing. Drawing a parallel to data creation on the traditional Web, most of which happens through Web forms, an analogous method to create data is needed on the semantic Web. An abundance of tools exist to support developers in creating such Web forms operating on a relational database scheme. Many of them also support the Model-View-Controller (MVC) pattern where a developer can generate scaffolding code (Web forms) that can be used to create, read, update and delete database entries based on the initial schema. To create such a Web form-based tool that operates based on an ontological schema, a number of challenges have to be addressed:

- Web form data is encoded in a key/value pair model (application/x-www-form-urlencoded) that is not directly compatible to the triple model of RDF. Therefore, a data binding mechanism is needed that binds the user input in Web form elements to an RDF model.
- Whereas a Web form based on a relational table has a fixed set of input fields based on the number of table columns, the RDF model is graph based with potential cycles. Further, RDF(s) properties are propagated from multiple superclasses (including inheritance cycles) and the types of properties for a class are not constrained by the definition (Open World assumption). Consequently, methods are required to decide on the properties to be displayed in a Web form for a given RDF node.
- In contrast to the relational model where tuples are bound to a relation (table), class membership for individuals in RDF(s) is not constrained for a class. Thus individuals that have been created as a type of a specific class need to be made available for reuse within a different class instance creation process.
- Beyond the standard datatypes in the relational model that can be easily mapped to different form input elements (e.g. String/Integer to text boxes, Boolean to radio buttons, etc.), the OWL model supports object properties that link individuals to other individuals via URIs. Object properties can also span multiple nodes in an RDF graph, forming a

property chain, i.e. they can refer to a class that is linked to another class through more than one property. To aid users in the creation of object properties methods have to be established to identify and link to existing individuals and to enable the creation of new individuals in the process of creating the object property.
- Once the Linked Data is created, access methods have to be defined that allow to retrieve the data with the Web form they have been created with, but also with any other Web form that supports the editing of the same types of relations that are defined in the RDF instance data. Further, these Web forms need to allow the user to add new properties to easily extend the RDF instance with new relations.
- Data creation in Web forms is often dependent on computational analysis functionality that cannot be expressed in RDF/OWL directly. Thus, these computations are implemented as services that are called after the completion of one Web form, and potentially lead to the creation of further Web forms based on the input data of the preceding Web form. This workflow of connecting Web forms for the creation and maintenance of Linked Data need to be captured in a model that can be accessed together with the Web form and RDF data model at each step in the process.

Some work [5, 4, 1, 7] exist that address some of the challenges above, but a model and tool that addresses the whole lifecycle of Linked Data creation and maintenance, including the ability to execute an explicit process model that controls the Linked Data lifecycle is still missing. Such a RESTful Linked Data workflow engine will support ordinary Web users in the process of creating Linked Data, eventually fulfilling one of the vision of the Semantic Web to provide a platform in which the same data is not created many times over again, but reused many times in different contexts.

References

1 S. Battle, D. Wood, J. Leigh, and L. Ruth. The Callimachus Project: RDFa as a Web Template Language, 2012.
2 J. Baumeister, J. Reutelshoefer, and F. Puppe. KnowWE: a Semantic Wiki for knowledge engineering. *Applied Intelligence*, 35:323–344, 2011.
3 S. Corlosquet, R. Delbru, T. Clark, A. Polleres, and S. Decker. Produce and consume linked data with drupal! In *Proceedings of the ISWC 2009*, pages 763–778. 2009.
4 A. Haller, T. Groza, and F. Rosenberg. Interacting with Linked Data via Semantically Annotated Widgets. In *Proceedings of JIST2011*, pages 300–317, 2011.
5 A. Haller, J. Umbrich, and M. Hausenblas. RaUL: RDFa User Interface Language – A data processing model for Web applications. In *Proceedings of WISE2010*, 2010.
6 A. Passant, J. G. Breslin, and S. Decker. Open, distributed and semantic microblogging with smob. In *Proceedings of the ICWE 2010*, pages 494–497. 2010.
7 S. Stadtmüller, S. Speiser, A. Harth, and R. Studer. Data-Fu: A Language and an Interpreter for Interaction with Read/Write Linked Data. In *Proceedings of WWW2013*, Rio de Janerio, Brasil, 2013.

3.7 On-the-fly Integration of Static and Dynamic Linked Data

Andreas Harth (KIT – Karlsruhe Institute of Technology, DE)

Joint work of Andreas Harth, Craig Knoblock, Kai-Uwe Sattler, Rudi Studer

The relevance of many types of data perishes or degrades over time. Consider the scenario where a user wants to catch a bus to go to a sporting event: the current location of the bus is more relevant than yesterday's or last week's position. Having access to such real-time information facilitates fast decision-making. However, the manual alignment of the information with additional sources to gain a deeper understanding of the tracked objects and the observed area is a labor-intensive process. The expending of up-front effort to access data in real-time may out-weight the advantage of having real-time information available.

To facilitate the integration of real-time sources, we propose to use a uniform approach to describe, access, and integrate real-time dynamic data, which in turn supports the integration with static geographical data and background knowledge. The types of supported data sources include:

- Static sources such as 2D maps, 3D models and point-of-interest (POI) data from XML files, Linked Data, and web APIs from the open web.
- Dynamic sources producing state updates (e.g. of moving objects), event information, etc. continuously or periodically, possibly with additional spatial/temporal properties, through web APIs.

A uniform approach enables the use of the integrated data in a variety of decision-supporting applications. For example, a real-time visualization that is annotated with background information can help a human user to make a well-informed and timely judgment of a situation. Additionally, software can automatically detect complex events based on the data and trigger appropriate notifications or actions. For example, a trigger could notify the user when to head for the bus stop based on the location of the approaching bus. Furthermore, having a uniform approach to describing, accessing, and integrating data sources enables the rapid discovery and addition of new relevant data sources. For example, the use of an existing bus application in a new country could require identifying and integrating new data sources.

Special value can be gained by integrating relevant data openly available on the web that has a quality and level of detail that is not achievable by a single, proprietary entity. Examples include data from OpenStreetMap or Foursquare that contain extensive geographical and meta-information about streets, buildings, businesses and other general points-of-interest (POIs). These data sets are maintained by huge numbers of contributing users.

To realize our goal, we need technologies that enable flexible, dynamic and scalable interactions of computing services and data sources.

As a key enabler for building such technologies we employ a uniform abstraction for components (resources) in large information systems termed Linked APIs. Components following the Linked API abstraction provide a standardized small set of supported operations and a uniform interface for both their data payload and their fault handling. Such a common abstraction allowing the manipulation of states as primitives enables us to specify the interactions between components declaratively both on the operational and data levels. We envision supporting the interoperability between web resources on an operational level similar to how semantic web technologies, such as RDF and OWL, support interoperability on a data level.

Given the minimal and unified interfaces of Linked APIs, we can use declarative means to specify interactions between different components in complex information systems [1]. The state manipulation abstraction and the high-level specifications describing the interplay between resources bring the following major benefits:

- Scalable execution: declarative specifications can be automatically parallelized more easily than imperative programs.
- Uniform and consistent error handling: instead of being confronted with source-specific error messages, universal error handlers can be realized.
- Substitution of resources: replacing a resource only requires adapting to the new vocabulary, while both the data model and the supported operations stay the same. Such flexibility is required as in large distributed information systems the underlying base resources may become unavailable and thus may put the entire networked application at risk.
- More flexible and cleaner specifications of interactions: the specifications can concentrate on the business logic of the intended interaction, while the operational interaction between components can be automated due to their standardized interfaces.

The overall benefits of such a method and apparatus are as follows:

- We may achieve real-time access to data integrated from several sources, some of them static and some of them dynamic.
- We can quickly integrate new data sources, as we use standard software interfaces to poll the current state of resources at specified time intervals or receiving updates and reacting on them, easing the transition from static to dynamic sources.
- We can quickly integrate new data sources, as the relation to the existing sources is specified declaratively, which allows for a high-level description of the interplay between resources that can be operationalized and optimized.

We believe that current web architecture offers the right abstraction and allows for the cost-effective implementation of such systems. Linked Data already allows for the integration of static data and the same mechanism can be used to achieve real-time functionality. In order to support interactive access to data, it will be necessary to execute extract/transform/load pipelines for performing integration at query time within seconds. We will also need tools to support end-users [2] in modeling real-time data sources to reduce the time needed in include a new live source into a constellation of systems that interoperate.

References

1 S. Stadtmüller, S. Speiser, A. Harth, and R. Studer. *Data-Fu: A Language and an Interpreter for Interaction with Read/Write Linked Data*. Proceedings of the 22nd International Conference on World Wide Web, WWW, pages 1225–1236, 2013.

2 M. Taheriyan, C. A. Knoblock, P. Szekely, and J. L. Ambite. *Rapidly Integrating Services into the Linked Data Cloud*. Proceedings of the 11th International Semantic Web Conference, ISWC, 2012.

3.8 Specifying, Executing, and Refining Complex Data Analytics Processes over Massive Amounts of Data

Melanie Herschel (University of Paris South XI, FR)

Overview

We are currently experiencing an unprecedented data deluge, as data in various formats is produced in vast amounts at an unprecedented rate. These properties of *variety*, *volume*, and *velocity* are at the core of the recent *Big Data* trend. One of the main issues of Big Data is how to make sense of the data that can no longer be handled or interpreted by a human alone. This commonly requires *integrating data* (to get a broader view of a subject) and then performing *analytical processes* on top of the data (to extract key figures of interest). In this context, we are particularly interested in how to specify and execute such Big Data Analytics processes. Section 3.8 presents our current efforts in this domain in the context of the Datalyse project.

Another observation is that in the dynamic context of Big Data, complex data transformation processes can no longer be designed and deployed once and left as-is afterwards, as has been the assumption for instance for data integration processes that require the design of a fixed global schema for the integrated result and the definition of a complex data integration workflow leading to the desired result. This observation has recently lead to the *pay-as-you-go* paradigm, for instance proposed for data integration in *dataspaces* [5]. The main idea is to get first results fast by a rapidly developed solution that is good enough to get started, to then refine the process subsequently. Section 3.8 describes our vision of supporting the evolution of data transformation processes throughout their complete lifecycle.

Big Data Analytics

The *Datalyse* project[3] on Big Data Analytics in the Cloud, started on May 1st 2013 and is a collaboration of several French industrial and academic partners (including Eolas, Business & Decision, INRIA, LIFL, LIG, and LIRMM). One objective is to provide a platform to facilitate developers to specify data analytical tasks over massive amounts of heterogeneous data provided by multiple data sources. As such, the platform provides both data integration and data analysis primitives that a developer can leverage when specifying the complete process in a declarative language. This language is compiled to be executed efficiently on a cloud based platform. In this context, we are particularly interested in the following aspects:

Data model. To be capable to provide processing primitives for a large variety of data sources, we first aim at defining a data model that captures this high variability. In particular, it should encompass NoSQL data models (such as JSON, XML, RDF) as well as associated schema specifications (e.g., XML Schema, RDFS). This data model, which extends previous work [4], will be used by processing primitives to access and manipulate data. These primitives include data access primitives, data transformation primitives (e.g., join, linking, aggregation) and data analytics primitives (e.g., data mining, clustering).

[3] See http://www.datalyse.fr/ for more details. The project is funded by the *Programme d'Etat des Investissements d'Avenir, Développement de l'Economie Numérique, Appel à projets "Cloud Computing" no. 3 – Big Data.*

Big Join and Linking. We will contribute to the definition, optimization and execution of both join and linking primitives. More precisely, we will first focus on *Big Join* (BJ), extending work of [8] to further take into account multi-source, heterogeneous, and continuous data. After developing an algebra for BJ manipulating data in the previously defined data model, we plan to develop and optimize BJ algorithms running on a Map/Reduce platform. We will follow a similar methodology for the *linking* primitive. The task of linking is very similar to a join, however, the goal is not only to join equal entries, but entries that refer to the same real-world entity, albeit being represented differently. Scalable linking has mainly focused on relational and hierarchical data [3] and we plan on further investigating scalable linking for complex data, for which few approaches have been proposed so far [1, 7].

Storing and indexing. The primitives described above are first defined at the logical level, before they are compiled into a physical plan that will be executed on a Map/Reduce platform. One essential aspect of physical plan execution is how to efficiently store and retrieve data, as demonstrated in [2]. To this end, we will investigate cloud storage and indexing strategies for data represented in our proposed data model and how to perform automatic storage optimizations.

Declarative Language. We will design a declarative language that allows developers to specify their analytical processes. This language then compiles into a logical plan using the primitives we consider. During compilation, we will perform optimizations targeted towards reducing the overall runtime of the process.

The developed techniques will be deployed in three real-world testbeds from different domains, i.e., monitoring, Open Government Data, and retail. In the first domain, we consider two use cases to ensure traceability, reporting, optimization, and analysis of irregular behavior w.r.t. energetic efficiency and IP network security, respectively. Concerning Open Data, we plan two use cases, i.e., one for data dissemination and one for data valorization. Finally, one retail use-case will focus on in-store and real-time business intelligence, whereas a second will concentrate on enriching catalogue data with semantic annotations.

Transformation Lifecycle Management

When developing complex data transformations, e.g., in the context of pay-as- you-go data integration, the data integration process (i.e., the data transformation leading to the integrated result) is gradually adapted and refined. The goal of transformation lifecycle management (TLM) is to semantically guide this refinement. The three main phases of TLM are (i) the *analysis* phase, where a developer verifies and analyzes the semantics of the data transformation (e.g., for debugging or what-if analysis), (ii) the *adapt* phase, where the transformation is changed (e.g., to fix a bug or to adapt to changing user requirements), and (iii) the *test* phase that helps in monitoring the impact of performed changes (e.g., to validate that the bug fix was indeed effective and no further error appeared).

Within the *Nautilus* project[4], we are devising algorithms and tools to semi-automatically support all phases of TLM. Currently, we are leveraging data provenance techniques for the analysis phase [6] of data transformations specified in a subset of SQL. We have also recently started investigating what query modifications can reasonably be suggested for SQL queries and how to compute a set of reasonable query modifications. All proposed solutions still

[4] http://nautilus-system.org/

lack efficient and scalable implementations, one avenue for future research. We also plan to support other data models and query languages in the future.

In the context of the *OakSaD* collaboration between the Inria Oak team and the database group at UC San Diego (`https://team.inria.fr/oak/oaksad/`), we plan to address the issue of analysis of complex business processes specified as data-centric workflows.

References

1 C. Böhm, G. de Melo, F. Naumann, and G. Weikum. Linda: distributed web-of-data-scale entity matching. In *International Conference on Information and Knowledge Management (CIKM)*, pages 2104–2108, 2012.

2 J. Camacho-Rodríguez, D. Colazzo, and I. Manolescu. Web Data Indexing in the Cloud: Efficiency and Cost Reductions. In *EDBT – International Conference on Extending Database Technology*, Genoa, Italy, Mar. 2013.

3 P. Christen. *Data Matching – Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection*. Data-centric systems and applications. Springer, 2012.

4 F. Goasdoué, K. Karanasos, Y. Katsis, J. Leblay, I. Manolescu, and S. Zampetakis. Growing Triples on Trees: an XML-RDF Hybrid Model for Annotated Documents. *VLDB Journal*, 2013.

5 A. Y. Halevy, M. J. Franklin, and D. Maier. Principles of dataspace systems. In *Symposium on Principles of Database Systems (PODS)*, pages 1–9, 2006.

6 M. Herschel and H. Eichelberger. The Nautilus Analyzer: understanding and debugging data transformations. In *International Conference on Information and Knowledge Management (CIKM)*, pages 2731–2733, 2012.

7 M. Herschel, F. Naumann, S. Szott, and M. Taubert. Scalable iterative graph duplicate detection. *IEEE Transactions on Knowledge and Data Engineering*, 24(11):2094–2108, 2012.

8 A. Okcan and M. Riedewald. Processing theta-joins using mapreduce. In *International Conference on the Management of Data (SIGMOD)*, pages 949–960, 2011.

3.9 Towards Vertebrate Linked Datasets

Aidan Hogan (National University of Ireland – Galway, IE)

The Problem...

Compared to the closed, cold, metallic, top-down and decidedly inorganic structure of a typical relational database, when I think of Linked Datasets, I think of evolving, seething, organic creatures with an emergent and complex structure. These organic creatures come in different shapes and sizes and can sometimes work together for various complementary purposes: a loosely interlinked ecology of sorts.

When I think of a Linked Dataset *like* DBpedia[5], I picture a gelatinous invertebrate entity—something like the creature from the 1958 movie "The Blob": organic, vaguely formed, turbid, expansive, expanding, wandering around in an Open World and occasionally assimilating some of the more interesting scenery. There is a certain unique beauty to this complex creature, no doubt, but it is a beauty that few can appreciate. If one is versed

[5] ... here selecting a prominent example purely for argument's sake.

in SPARQL, one can strike up a conversation with the creature and ask it a variety of questions about the things it has assimilated. If you phrase your question right (use the correct predicates and so forth) and the creature is in the mood (the endpoint is available), and it has assimilated the things you're interested in (has the data), you can sometimes get a response.

But it is an enigmatic creature: oftentimes the response will not be what you expected. For example, if you ask DBpedia something simple like how many countries are there:

```
SELECT (COUNT(?c) as ?count) WHERE { ?c a dbpedia-owl:Country }
```

it tells you 2,710. Sounds a bit high, no? Should be just shy of 200? Why did DBpedia say 2,710?

The Position...

Let's leave the admittedly belaboured and entirely unfair analogy of The Blob aside.[6] Linked Datasets are often analogous to black boxes. All we typically know about these black boxes are the the shape of content they house (triples), the size of that content (number of triples), the category of that content (domain of data), and some features of that content (classes and properties used).[7] Figuring out the rest is left as an exercise for the consumer. Oftentimes this is simply not enough and the consumer loses patience. So the question then is: how best to describe the contents of these black boxes.

The main aim of such a description should be to inform a consumer as to what expectations of freshness, accuracy and coverage of results they can expect for various types of queries over various chunks of the data (and let's assume that the endpoint performs perfectly now; an assumption that does not nearly hold in practice). The description should summarise the content of the black-box in such a way that it can be searched and browsed as a catalogue: it should allow a consumer to, hopefully at a few glances, *make sense* of the dataset they are querying, or at least a *core* of the dataset. My position in a few words: *we need intuitive mechanisms for humans to make sense of Linked Datasets (in whole and in part).*

How can this be done? Relying on the semantics of the used terminology, as given in a vocabulary/ontology, says nothing about the data itself. We can look at instances of classes and common relationships between instances of various classes, which is quite a useful exercise, but class- centric descriptions do not tell much of the story. Again, in DBpedia, we have 2,710 instances of `Country`. So what definition of `Country` permits 2,710 instances? We can play around a bit and determine that a lot of listed countries no longer exist, or are not independent states, or are aliases or historical names, etc. But which are the current countries? And how many countries have flags or populations defined? How recent are those population measures?

It's great that DBpedia and other such datasets have lots of organic tidbits of information like historical countries and aliases and so forth. However, the result is a highly "non-normalised" soup of data that is difficult to describe and difficult to query over. We need a little *ordo ab chao*. One solution would be to trim the fat from datasets like DBpedia so they fit into a highly normalised database-like schema that best fits the most common needs and that gives consumers a smoother experience. But this is not only unnecessary, it is inorganic, inflexible and, dare I say it, not very Semantic-Webby.

---

[6] I have nothing but cautious affection for DBpedia and Linked Data.

[7] If you're very lucky, you might find all of these details encapsulated in a VoID description.

One Proposal...

Instead of imposing a rigid inorganic structure on Linked Datasets like DBpedia so that they fit neatly into familiar rectangular frames of conceptualisation, perhaps we can just try find a natural shape to the dataset: a "spine". We can begin by clustering instances in an extensional sense: looking at clusters of instances defined using the same predicates and the same types. For example, a cluster might be a group of instances that all have type `Country`, at least one `capital`, at least one `gdp`, and at least one `city`. We can call these clusters "abstract classes" or "prototypes" or "templates" or "least common factors" or "instance cores" or ... well in fact, such clusters are not even an entirely new idea (and are similar to, e.g., "characteristic sets" [2]) but no matter.

More important is what we can use these clusters for. In this model, clusters naturally form a subsumption hierarchy where instances within a cluster are also contained within less specific sub-clusters. The number of clusters and the level of detail they capture can be parameterised for their computation [1]. A person can browse the hierarchy of clusters– -the "spine"—of a dataset to see at a glance what it contains and to find the cluster he/she can target with a query. One might start exploring the super-cluster containing instances with type `Country` and see it branch into one sub-cluster with 910 instances with `dbprop:dateEnd` (dissolved countries) and a disjoint sub-cluster of 191 instances with the subject `category:Member_states_of_the_United_Nations` (current countries recognised by the UN; 2 are missing). Browsing the hierarchy thus helps with understanding the scope and breadth of data in increasing detail, helps with disambiguation, and helps with formulating a query that targets only the instances of interest.

Furthermore, the richer and more complete the clusters, the higher the degree of homogeneity in those instances. Child-clusters in the hierarchy with similar cardinalities may indicate incomplete data: e.g., taking the UN member cluster of 191 instances, we find a sub-cluster with of 188 instances with defined capitals, where we could conclude that capitals are missing in 3 instances. Filling in the blanks will merge one step of the hierarchy and increase the homogeneity of the country descriptions. Merging highly-overlapping sub-clusters in the hierarchy then becomes a quantifiable bottom-up goal for local normalisation.

Subsequently, clusters can be annotated for the instances they encapsulate; e.g., in this cluster, capitals rarely change, populations are all as recent as 2008, GDP values are 87% accurate, etc. A directed (subject- to-object), labelled (predicate), weighted (count) graph can be constructed between clusters as the aggregation of the most common links between their instances. Furthermore, the integration of two or more Linked Datasets can then be coordinated through these clusters, with the goal of identifying and consolidating conceptually overlapping clusters to a high (and easily quantifiable) degree.

The resulting spine of the dataset then gives a core and a shape to the dataset; an entry point to follow; a way of distinguishing normalised and complete data from non-normalised and incomplete data; a basis for coordinating integration; an emergent structure from which all the other organic matter can extend.

References

1 J. H. Friedman and J. J. Meulman. Clustering objects on subsets of attributes. *J. R. Stat. Soc*, 66:815–849, 2004.

2 T. Neumann and G. Moerkotte. Characteristic sets: Accurate cardinality estimation for RDF queries with multiple joins. In *ICDE*, pages 984–994, 2011.

3.10 Interoperability for Linked Open Data and Beyond

Katja Hose (Aalborg University, DK)

Querying Linked Open Data

During the past couple of years, query processing as a foundation to achieve interoperability has been a very active field of research in the Semantic Web community. Especially, efficiently processing SPARQL queries over RDF data in general and Linked Open Data in particular attracted attention [9, 7, 5, 1, 2] with some approaches being conceptually closer to data integration in database systems than others.

Assuming a number of independent Linked Data sources (SPARQL endpoints), we have designed a framework and optimized distributed query processing of SPARQL queries over RDF data [10]. In addition to generating efficient query execution plans involving efficient implementations of operators, another problem in distributed systems is source selection, i.e., the decision whether a source should be considered during query processing or not. Apart from basic approaches such as budgets or time-to-live constraints, indexes are a key ingredient to enable source selection. On the one hand, we have developed approaches that estimate how many results a source might contribute to (part of) the query [4]. On the other hand, we have also considered benefit-based source selection [6] that also considers the overlap in the data of available sources and estimates the benefit of querying a particular source in terms of the unique new query results its data will produce. To make local processing at a single source with large amounts of RDF data more efficient, we have also been working on providing efficient solutions to SPARQL query processing in parallel scale-out systems by splitting the data into partitions and assigning the partitions to cluster nodes in a certain way that we can exploit during query processing [3].

Collaborative Knowledge Networks

The works mentioned so far focus on efficient processing of structured queries (SPARQL) over RDF sources, which is only one problem of many to achieve interoperability. Therefore, we have recently proposed a framework, Colledge [8], with the vision to support a higher degree of interoperability by allowing for interaction and combining aspects of P2P systems, social networks, and the Semantic Web.

In Colledge, functionalities such as query precessing, reasoning, caching, data sources, wrappers, crowd sourcing, information extraction, etc. are modeled as services offered by nodes participating in the network. Queries are processed in a P2P-like manner and typically involve chains of nodes that are automatically selected and not known a priori. The user provides feedback on the correctness and relevance of the results, which is propagated based on collected provenance information to the nodes involved. This process helps detecting inconsistencies and errors, improves the query result over time, and eventually leads to updates of the original data, hence allowing for collaborative knowledge that is developing over time.

Challenges

In general, the problem of overcoming heterogeneity naturally arises whenever we want to enable interoperability between multiple data sources. This problem is not new and has

played a prominent role in research for quite some time now. Regarding interoperability for RDF and Linked Open Data sources alone, there are still a number of open issues, especially taking efficiency, updates, and reasoning into account. But there is also a great potential to learn from the advances and mistakes of other communities.

However, what we are mostly doing is developing solutions designed for a particular (sub)community, a bit of mapping and wrapping here and there. The question that remains to be answered is: Is this enough to handle the huge degree of heterogeneity that is coming along with accessing data on the Web (relational data, SQL, XML, XQuery, HTML, RDF, Linked Open Data, SPARQL, reasoning, information extraction, ontology matching, updates, Web services,...)?

Another questions that arises naturally is whether such a complex system is really what we need and want in the future. And if we want it, what are the main challenges? And what would be the best way to achieve interoperability then, (i) building upon the "old" ways and extending them to support the new problems that are introduced or (ii) is it possible to approach the whole problem in a different, novel way?

References

1 Carlos Buil Aranda, Marcelo Arenas, and Óscar Corcho. Semantics and optimization of the SPARQL 1.1 federation extension. In *ESWC (2)*, pages 1–15, 2011.
2 Valeria Fionda, Claudio Gutierrez, and Giuseppe Pirrò. Semantic navigation on the web of data: specification of routes, web fragments and actions. In *WWW*, pages 281–290, 2012.
3 Luis Galarraga, Katja Hose, and Ralf Schenkel. Partout: A distributed engine for efficient rdf processing. *CoRR*, abs/1212.5636, 2012.
4 A. Harth, K. Hose, M. Karnstedt, A. Polleres, K., and J. Umbrich. Data summaries for on-demand queries over linked data. In *WWW*, pages 411–420, 2010.
5 Olaf Hartig. Zero-knowledge query planning for an iterator implementation of link traversal based query execution. In *ESWC (1)*, pages 154–169, 2011.
6 Katja Hose and Ralf Schenkel. Towards benefit-based rdf source selection for sparql queries. In *SWIM*, pages 2:1–2:8, 2012.
7 Andreas Langegger, Wolfram Wöß, and Martin Blöchl. A Semantic Web middleware for virtual data integration on the Web. In *ESWC*, pages 493–507, 2008.
8 Steffen Metzger, Katja Hose, and Ralf Schenkel. Colledge – a vision of collaborative knowledge networks. In *2nd International Workshop on Semantic Search over the Web (SSW 2012)*, page . ACM, 2012.
9 Bastian Quilitz and Ulf Leser. Querying distributed RDF data sources with SPARQL. In *ESWC*, pages 524–538, 2008.
10 A. Schwarte, P. Haase, K. Hose, R. Schenkel, and M. Schmidt. FedX: Optimization techniques for federated query processing on linked data. In *ISWC*, pages 601–616, 2011.

3.11 A Process-Oriented View of Website Mediated Functionalities

Martin Junghans (KIT – Karlsruhe Institute of Technology, DE)

Functionalities that are offered in form of interactive websites are omnipresent. It requires a substantial manual and tedious effort to use the functionalities within increasingly sophisticated use cases. Structured descriptions enable the development of methods that support users in dealing with everyday tasks. We present how a process-oriented view of the Web is

gained and helps end users, for instance, in finding information scattered across multiple websites.

The Web is not only an interlinked collection of documents. Functionalities provided in form of interactive Web pages and Web applications are offered and consumed on a regular basis by most of the end users connected to the Internet. Web pages serve as a front-end to access services and information of the Deep Web. In contrast to the Semantic Web with its aim to allow providers to annotate their services such that automatic discovery and composition are enabled, website mediated functionalities target primarily at human use.

By observing end users and their **browsing behavior**, the Web is perceived as a pool of functionalities solving simple tasks. Users select functionalities and use them in a certain sequence in order achieve a goal. For example, in order to arrange a travel or to buy a high rated and cheap product, several individual functionalities are composed in the users' minds. Logic dependencies between inputs and outputs have to be managed manually and often the same inputs are provided at multiple websites repeatedly.

A large portion of the Web can be seen as a set of distributed and networked processes. They can provide access to information and cause effects during their execution. Also, they can require multiple interactions (such as form submissions and link selections) with the user. Unfortunately, these processes are currently not explicit. Users have to compose them every time again on their own. However, a **process-oriented view** on the Web requires an explicit and structured description of the functionalities and processes. In our approach, we model end user browsing processes that describe how users have to interact with which website at which time in order to consume functionalities and reach the desired outcomes. The descriptions allow to support users in combining, sharing, and reusing the processes, which capture previous efforts to achieve a goal [1].

We learned from the lack of public Semantic Web Services that we cannot rely on providers to create semantic annotations and to adopt a top-down formal semantics of Web services. So, we let users capture their browsing processes by existing **Web automation scripting** tools, which monitor, describe, and partially automate the process execution. In a bottom-up approach [2], semantic annotation can be added by users when needed, e.g., to describe elements of interactions and Web pages. Semantic annotations of websites can also be derived from the scripts [3].

Information Search Based on a Process-Oriented View

For many practical purposes, end users need information that is scattered across multiple websites. Consider for example an end user who is interested in knowing the names of the chairs of a particular track at the previous WWW conferences. As of today, Web search engines do not deliver satisfactory results for queries similar to "track chairs of all WWW conferences". In order to obtain the required information the end user has to pose multiple queries to a search engine, browse through the hits, and aggregate the required information fragments outside of the found Web pages.

End users need help in selecting the pages that are relevant for obtaining the scattered information. Such a help must contain at least the set of the pages that the end user should visit, and support for invoking all the pages of the set easily. More advanced help could contain the complete end user browsing process including support for data flow between the user and the pages as well as among the pages, and control flow if there are data dependencies among inputs and outputs of Web pages in the set. We aim at providing the end users with a list of browsing processes that are relevant for a given information need instead of a list of links to Web pages. Each browsing process in the list of hits will lead the end user to the required information.

In order to search for existing browsing processes, e.g., from a repository shared with other users, we developed efficient **discovery** techniques. We proposed an offline classification of processes [4], which is based on formally defined classes, and the use of offline and online index structures [5] to efficiently locate desired browsing processes from large repositories. The discovery allows to reuse existing browsing efforts. The composition of browsing processes promises to create solutions to the information need that have not been executed before.

Acknowledgments

This paper presents results of joint efforts of a group of colleagues including Sudhir Agarwal (Stanford University), Charles J. Petrie (Stanford University), and the author.

References

1 Agarwal, S., Junghans, M.: *Swapping out coordination of web processes to the web browser.* Brogi, A., Pautasso, C., Papadopoulos, G. (eds.), ECOWS, IEEE pages 115–122, 2010.
2 Agarwal, S., Petrie, C.J.: *An alternative to the top-down semantic web of services.* IEEE Internet Computing 16(5), pages 94–97, 2012.
3 Agarwal, S., Junghans, M.: *Towards simulation-based similarity of end user browsing processes.* Daniel, F., Dolog, P., Li, Q. (eds.), ICWE, pages216–223, Springer, 2013.
4 Junghans, M., Agarwal, S., Studer, R.: *Behavior classes for specification and search of complex services and processes.* Goble, C., Chen, P., Zhang, J. (eds.), ICWS, pages 343–350, IEEE, 2012.
5 Junghans, M., Agarwal, S.: *Efficient search for web browsing recipes.* ICWS, IEEE (to appear), 2013.

3.12 Merits of Hypermedia Systems

Kjetil Kjernsmo (University of Oslo, NO)

Research Interests

My primary research interest is the optimisation of SPARQL queries in a federated regime, as we have noted that this is not practical because the federation engine has insufficient information to optimise, or the information is so large that it defeats the purpose of optimisations to begin with. I plan to help remedy this problem by computing very compact digests and expose them in the service description. I have not yet published any articles on this topic, but the research is in the immediate extension of SPLENDID[3]. My secondary research interest is using statistical design of experiment in software performance evaluation.

However, coming from an industry background in software development, experience suggests that the above research interests does not adequately address many immediate needs when developing information systems to process the rapidly increasing amount of available data. I believe that large SPARQL-driven systems would be the "right tool for the job" in only a limited, and currently unclear, set of cases. Further exposure to new ideas in the developer community lead me to develop a third interest, namely hypermedia RDF.

Problems with SPARQL interfaces are many: They require extensive training of developers; it is not immediately clear what data are available and what may done with the data; it is easy to formulate queries that will cause the endpoint to become overloaded and hard to

protect against them without also rejecting legitimate queries; it places heavier systems in the execution path of an application, etc.

Hypermedia RDF

In [4], I examined some practical implications of the HATEOAS constraint of the REST architectural style, see [2] Chapter 5, and in that light argued why hypermedia RDF is a practical necessity.

Mike Amundsen defines hypermedia types[1] as

> Hypermedia Types are MIME media types that contain native hyper-linking semantics that induce application flow. For example, HTML is a hypermedia type; XML is not.

We continued to derive a powerful hypermedia type based on RDF within a classification suggested by Mike Amundsen. Since this publication, I have also noted that the "embedded links" factor can be achieved by using `data` URIs, thus satisfying all but one of the factors proposed by Mike Amundsen.

Further, we noted that other interesting factors is the self-description, which is a important characteristic of the RDF model, and other minor concerns.

To bring forward a concrete example of how to make a serialised RDF graph into a hypermedia type, we suggest adding some triples to every resource (where prefixes are omitted for brevity):

```
<> hm:canBe hm:mergedInto, hm:replaced, hm:deleted ;
   hm:inCollection <../> ;
   void:inDataset [void:sparqlEndpoint </sparql> .] .
```

Possible Uses

Over time, I believe that interfaces that require developers to read external documentation will loose to interfaces where "View Source" is sufficient to learn everything that is needed. This is the essence of hypermedia systems, in the RDF case, everything needed to program is available in the RDF. It tells application what it may do next.

In the above, we have included mostly create, add and delete primitives, but these could be refined for application scenarios if needed.

For example a pizza baker publishes Linked Data about pizzas they sell, including data sufficient to create a sophisticated search and sales application. The Linked Data will then include triples that state explicitly how the order should be placed, i.e. what resources need updating. Moreover, the may not only want to sell pizzas, but also drinks. Thus, the Linked Data presented to the application should not only be a Symmetric Concise Bounded Description, as common today, but careful designed to provide exactly the data needed to optimise sales.

References

1 Mike Amundsen. Hypermedia Types. http://amundsen.com/hypermedia/, 2010.
2 Roy Thomas Fielding. *Architectural Styles and the Design of Network-based Software Architectures*. PhD thesis, University of California, Irvine, 2000.
3 Olaf Görlitz and Steffen Staab. SPLENDID: SPARQL Endpoint Federation Exploiting VOID Descriptions. In *Proceedings of the 2nd International Workshop on Consuming Linked Data*, Bonn, Germany, 2011.

4 Kjetil Kjernsmo. The necessity of hypermedia rdf and an approach to achieve it. In *Proceedings of the First Linked APIs workshop at the Ninth Extended Semantic Web Conference*, May 2012.

3.13 Next Generation Data Integration for the Life Sciences

Ulf Leser (HU Berlin, DE)

Joint work of Sarah Cohen-Boulakia, Ulf Leser

Ever since the advent of high-throughput biology (e.g., the Human Genome Project), integrating the large number of diverse biological data sets has been considered as one of the most important tasks for advancement in the biological sciences. Whereas the early days of research in this area were dominated by virtual integration systems (such as multi-/federated databases), the current predominantly used architecture uses materialization. Systems are built using ad-hoc techniques and a large amount of scripting. However, recent years have seen a shift in the understanding of what a "data integration system" actually should do, revitalizing research in this direction. We review the past and current state of data integration for the Life Sciences and discuss recent trends in detail, which all pose challenges for the database community.

3.14 Service- and Quality-aware LOD; the Yin and Yang of Complex Information Ecosystems

Andrea Maurino (University of Milan-Bicocca, IT)

Introduction

In Chinese philosophy, the concept of yin and yang is used to describe how seemingly opposite or contrary forces are interconnected and interdependent in the natural world; and, how they give rise to each other as they interrelate to one another[8]. From a user perspective data and services are the yin and yang principle of complex information ecosystems. Data search and service invocation can be individually carried out, and both these operations provide value to users in the context of complex interactions. However, typically, users need to perform aggregated searches able to identify not only relevant data, but also services able to operate on them. The research on data integration and service discovery has involved from the beginning different (not always overlapping) communities. As a consequence, data and services are described with different models, and different techniques to retrieve data and services have been developed. Nevertheless, from a user perspective, the border between data and services is often not so definite. According to Chinese philosophy Yin and yang are actually complementary, not opposing, forces, interacting to form a whole greater than either separate part[1] and data and services provide a complementary view of the same ecosystems.

[8] http://en.wikipedia.org/wiki/Yin_and_yang

Data provide detailed information about specific needs, while services execute processes involving data and returning an informative result. The advent of the Web of data and in particular linked open data (LOD) that is open data semantically interconnected, could help the definition of new approaches to effectively access data and services in a unified manner due to the fact that LOD and semantic web services speak the same language. Strictly related to the previous issue, quality in LOD is another new, challenging and relevant issue. While there is broad literature on data quality (in particular on both relational or structural data [1] there are very few works that consider the quality of linked open data [**?**]. Quality in LOD is a crucial problems for the effective reuse of published data and recently it has increased its relevance in the context of the research communities due to the availability of existing data. At the ITIS lab of University of Milano Bicocca both problems has been considered and in this position paper I report the most important results and the main issues to be still solved.

Quality in Linked Open Data

In my vision the multiple (unconsciously) marriage of database, sematic web and web communities brought the birth of linked open data. By considered LOD principles and its applications, we quickly recognize that parents of LOD are the three communities; because data must to be modeled, queried, indexed and so on (classic database topics), data must to be semantically linked and it is published on the Web by publishing their URIs. As a consequence, LOD brings some old issues coming from existing parents, but there are also new ones coming from the marriage. Quality in LOD is a typical example. In database community data quality is very well studied and a lot of methodologies and dimensions has been defined [3], but in LOD new issues arise and older ones are different. For example let consider the time related dimensions (such as currency) it is an important quality dimension that can easily managed in relational database by means of log file and temporal db. In LOD domain time, related information are very important because they can be used as proxy for the evaluation of validity of a rdf triple, but as recently shown [4] current practices on LOD publishing does not include such metadata. Moreover the completeness dimension is easily defined and assessed in the database community thanks to the "closed world assumption", while at the level of Web this assumption is not correct and this make more and more difficult to measure the completeness of LOD. The study of quality in LOD is made harder due to the fact quality is in data consumer's eyes not data producer's ones and so techniques for assessing and improving LOD are strongly related to the need of data producers that are not known when data is published.

Service and Linked Open Data Integration

In [5] with other colleagues, I proposed a solution for aggregated search of data and services. We started by the assumption to have different and heterogeneous datasources and a list of semantic web services. In the proposed solution we first build the data ontology by means of the momis approach [6] then we create a service ontology by considering the semantic description of available web services. A set of mappings between the data and service ontology allow the possibility to search both data and services. A framework architecture comprising the data ontology (DO) manager, which supports the common knowledge extracted from heterogeneous sources, and XIRE, an information retrieval-based Web Service engine able to provide a list of ranked services according to a set of weighted keywords are designed and developed to evaluate the effectiveness of the overall approach. Evaluations based on

state-of-the-art benchmarks for semantic Web Service discovery shown that our information retrieval-based approach provides good results in terms of recall and precision. With the advent of quality enhanced LOD the proposed approach in [5] is still valid due to the fact that the data ontology can be considered as the LOD cloud that can be easily queried by SPARQL endpoints. Thanks to the fact that both data and services speaks the same (ontological) language the integration process can be easily considered as an ontology instance matching problems. The main issue in the integration of service and linked open data is the availability of semantically described services that it is still an open problem.

References

1 Sadiq, S., ed. *Handbook of Data Quality*. Springer, 2013.
2 Zaveri, A., Rula, A., Maurino, A., Pietrobon, R., Lehmann, J., Auer, S. *Quality assessment in linked open data*. Submitted for publication, 2013.
3 Batini, C., Cappiello, C., Francalanci, C., Maurino, A. *Methodologies for data quality assessment and improvement*. ACM Comput. Surv. 41(3), 2009.
4 Rula, A., Palmonari, M., Harth, A., Stadtmüller, S., Maurino, A. *On the diversity and availability of temporal information in linked open data*. Proc. International Semantic Web Conference, pages 492–507, Springer, 2012.
5 Palmonari, M., Sala, A., Maurino, A., Guerra, F., Pasi, G., Frisoni, G. *Aggregated search of data and services*. Inf. Syst. 36(2) pages 134–150, 2011.
6 Bergamaschi, S., Castano, S., Vincini, M., Beneventano, D. *Semantic integration of heterogeneous information sources*. Data Knowl. Eng. 36(3), pages 215–249, 2001.

3.15 A Purposeful View of Data: Getting the Data When You Need and How You Like It

Sheila McIlraith (University of Toronto, CA)

Most data, whether structured, semi-structured or unstructured, is acquired for a purpose, and that purpose often necessitates its composition with other data, and sometimes its transformation or aggregation via distinct sub-processes. Such composition and processing may be as simple as a set of join operations, typical in relational query processing, or as complex as a workflow of data analytics computations, selected on the fly, based on intermediate outcomes. In this talk, we look at the complex, evolving ecosystem of data and programs on the web and in the cloud, through the lens of (business) processes. We argue that (business) processes provide a vehicle for the customized and optimized selection, acquisition, integration, transformation, and display of data. Informally, the purpose for which the data is being collected can be described by a (business) process, mandating constraints on what data is collected, when it is collected and how it is transformed. I will report on our ongoing work in developing such optimized data-aware processes. This will be followed by a brief discussion of the relationship of this endeavor to the vision of Semantic Web Services, and a reflection on what, if anything, Semantic Web Services and the lessons learn from that effort have to contribute to the general task of developing complex interoperating data and sub-process ecosystems.

3.16 Position Statement

Bernhard Mitschang (Universität Stuttgart, DE)

The topic of the seminar is getting more and more important, since we do not only produce more and more data, we even do not erase data (we got somehow) anymore. Furthermore, the need to relate different portions of data is also getting more important due to informedness and actuality needs as well as for decision making needs. So the question naturally arises: What can/should we do from a technical perspective as well as from an economic perspective, and who pays the bill?

The technical view: all is there, somehow, see below. Money: well, if someone benefits from it, then, of course, this one pays, or it is free of charge, because of crowdsourcing, Google and the like, or governments pay for it.

I assume that to this seminar the first topic is the important one. So let me detail a bit on this: I can see that there is a bunch of ready-to-use and well understood technologies and systems that are perfectly suitable. For example:

1. ETL: transform and integrate starting from data portions taken from various data sources.
2. Data Analytics: aggregate, condense
3. Stream processing: "on the fly" processing of incoming data
4. Mashup technology, like Yahoo Pipes: ad hoc

I would like to discuss, whether these techniques are sufficient or not, and, if there is a need for new technology, what are the properties and characteristics.

3.17 Next Generation Data Profiling

Felix Naumann (Hasso-Plattner-Institut – Potsdam, DE)

Profiling data is an important and frequent activity of any IT professional and researcher. We can safely assume that any reader has engaged in the activity of data profiling, at least by eye-balling spreadsheets, database tables, XML files, etc. Possibly more advanced techniques were used, such as key-word-searching in data sets, sorting, writing structured queries, or even using dedicated data profiling tools. While the importance of data profiling is undoubtedly high, and while efficiently and effectively profiling is an enormously difficult challenge, it has yet to be established as a research area in its own right.

References

1 Felix Naumann. *Data Profiling Revisited.* In SIGMOD Record, 2013.

3.18 Bridging Real World and Data Spaces in Real-time: Data Stream Management and Complex Event Processing

Daniela Nicklas (Universität Oldenburg, DE)

Motivation

With the upcoming widespread availability of sensors, more and more applications depend on physical phenomena. Up-to-date real-world information is embedded in business processes, in production environments, or in mobile applications, where context-aware applications can adapt their behavior to the current situation of their user or environment. However, while more and more sensors observe the world, the ratio of data which is actually used is decreasing – we are drowning in a sea of raw data. Big data is often characterized in the dimension of volume, variety, and velocity. Dealing with this upcoming and ever-increasing stream of sensor data is not easy: it is often lowlevel (just raw sensor readings with no interpretation yet), distributed, noisy, bursty, and comes from heterogeneous sources, ranging from simple stationary single-value sensors (e.g., a thermometer) over mobile measurements to high-volume sensors like cameras or laser scanners. And archiving the data leads to ever-increasing storage needs. So, we face all three challenges of big data.

Background

Our approach to deal with these challenges is to develop generic data management systems for streaming data. The goal of data stream management is to provide the same flexibility and data independence for data streams as for stored data. For sensor data representing real-world situations, we need additional operators, e.g., to deal with different semantic layers (from raw data over features to objects/entities) or sensor data quality. Thus, we combine techniques from database management, probabilistic databases, sensor data fusion, and context-aware computing to create new base technology for these smart applications of the future. Our current application scenarios are highly dynamic world models for autonomous vehicles [5], safe offshore operations [6], ambient assisted living [4], and smart cities [2], and we implement our concept in the open source data stream framework Odysseus [1].

Open Questions

From this motivation (and from my background) we might discuss the following topics / challenges:

- Ecosystems of stream and non-stream data: how would system architectures look like that combine streaming and event-based data with large amounts of stored data? Could federated architectures be a solution [3]?
- Knowledge management: how can we manage supervised and un-supervised data mining techniques with online observation of relevant situations in streaming environments?
- Challenges from application domains: smart cities and/or smart factories (and the new German keyword industrie 4.0)

References

1 H.-Juergen Appelrath, Dennis Geesen, Marco Grawunder, Timo Michelsen und Daniela Nicklas. *Odysseus: a highly customizable framework for creating efficient event stream*

management systems. Proceedings of the 6th ACM International Conference on Distributed Event-Based Systems, DEBS '12, pages 367–368, 2012.

2 Marcus Behrendt, Mischa Böhm, Mustafa Caylak, Lena Eylert, Robert Friedrichs, Dennes Höting, Kamil Knefel, Timo Lottmann, Andreas Rehfeldt, Jens Runge, Sabrina-Cynthia Schnabel, Stephan Janssen, Daniela Nicklas und Michael Wurst. *Simulation einer Stadt zur Erzeugung virtueller Sensordaten für Smart City Anwendungen (Demo)*. 15. GI-Fachtagung Datenbanksysteme für Business, Technologie und Web, Magdeburg, 2013.

3 Andreas Behrend, Daniela Nicklas und Dieter Gawlick. *DBMS meets DSMS: Towards a Federated Solution*. Proceedings of 1st International Conference on Data Technologies and Applications (DATA'12). SciTePress, 2012.

4 Dennis Geesen, Melina Brell, Marco Grawunder, Daniela Nicklas und H.-Juergen Appelrath. *Data Stream Management in the AAL – Universal and Flexible Preprocessing of Continuous Sensor Data*. Reiner Wichert und Birgid Eberhardt (eds). Ambient Assisted Living, pages 213–228. Springer Verlag, 2012.

5 Christian Kuka, Andre Bolles, Alexander Funk, Sönke Eilers, Sören Schweigert, Sebastian Gerwinn und Daniela Nicklas. *SaLsA Streams: Dynamic Context Models for Autonomous Transport Vehicles based on Multi-Sensor Fusion*. Proceedings of IEEE MDM 2013, 14th International Conference on Mobile Data Management, Milan, Italy, 2013.

6 Nils Koppaetzky und Daniela Nicklas. *Towards a model-based approach for context-aware assistance systems in offshore operations*. Workshop proceedings of 11th Annual IEEE International Conference on Pervasive Computing and Communications, 2013.

3.19 Cartography on the Web

Giuseppe Pirrò (Free University of Bozen-Bolzano, IT)

Joint work of Valeria Fionda, Claudio Gutierrez, Giuseppe Pirrò

Cartography is the art of map making. Abstractly, it can be seen a set of information transformations aimed at reducing the characteristics of a large selected area and putting them in a visual image, that is, a map. A map should meet the property of *abstraction*, that is, it has to be always smaller than the region it portraits. Besides, its visual representation plays a fundamental role for facilitating its interpretation by a map (human) user.

When mapping physical landscapes there are well-consolidated cartographic principles and techniques. However, we live in the Web era and thus applying cartographic principles also to digital landscapes becomes intriguing. Similarly to the Earth, the Web is simply too large and its interrelations too complex for anyone to grasp much only by direct observation. The availability of maps of Web regions is fundamental for helping users to cope with the complexity of the Web. Users via Web maps can track, record and identify conceptual regions of information on the Web, for their own use, for sharing/exchanging with other users and/or for further processing (e.g., combination with other maps). Indeed, Web maps are useful to find routes toward destinations of interest, navigate within (new) complex domains, and discover previously unknown connections between knowledge items.

Toward the development of Web maps, there are some challenging research issues. First, the Web is huge, distributed and continuously changing; therefore, techniques to specify, access and retrieve parts of it, that is, regions of interest are needed. Second, given a region of the Web, what is a reasonable definition of map? Third, is it feasible to efficiently and automatically build Web maps? Besides these points that focus on how to project

the principles of traditional cartography to the Web, additional research challenges emerge. One is the possibility to go beyond maps for only human users. This sets the need for having a mathematical model of regions and maps of the Web so that their properties and mutual relationships can be rigorously defined and understood. Hence, maps can be given a machine-readable format, which will foster their exchange and reuse. Moreover an algebra for maps can be defined, which will enable their combination via well-defined operations like union and intersection.

Nowadays, tools like bookmarks and navigational histories touch the problem of building maps of the Web. However, they do not comply with the notion of map that we envision for several reasons. First, they do not meet the property of abstractions: the Web region itself, that is the set of pages visited or bookmarked, is the map. Second, the map is a set of (disconnected) points; relations among them are lost. Third, these approaches rely on the manual activity of the Web user and thus are not suitable for the automation of the process of building maps. Last but not least, they are designed only for final human consumption. Initiative like Topic Maps face the problem of standardizing information management and interchange. The focus here is to manually create visual representations; Topic Maps cannot be automatically constructed (from the Web) and do not include an abstraction phase.

We investigate the applicability of cartographic principles to the Web. We model the Web space as a graph where nodes are information sources representing objects (e.g., people) and edges links between them (e.g., friendship). With this reasoning, a region becomes a (connected) subgraph of the Web and a map a (connected) subgraph of a region. Hence, we formalize the general problem of obtaining different kinds of maps from a graph. Our formalization makes maps suitable to be manipulated via an algebra also defined. To automate the construction of maps, we define a general navigational language that differently from existing languages, returning sets of (disconnected) nodes, returns regions. Then, we devise algorithms to efficiently generate maps from these regions. We instantiate and implement our map framework over the Web of Linked Open Data where thousands of RDF data sources are interlinked together. The implementation along with examples is available online[9].

3.20 Completeness of RDF Data Sources

Giuseppe Pirrò (Free University of Bozen-Bolzano, IT)

Joint work of Fariz Darari, Werner Nutt, Giuseppe Pirrò, Simon Razniewski
Main reference F. Darari, W. Nutt, G. Pirrò, S. Razniewski, "Completeness Statements about RDF Data Sources and Their Use for Query Answering," in Proc. of the 12th International Semantic Web Conference, LNCS, Vol. 8218, pp. 66–83, Springer, 2013.
URL http://dx.doi.org/10.1007/978-3-642-41335-3_5

With thousands of RDF data sources today available on the Web, covering disparate and possibly overlapping knowledge domains, the problem of providing high-level descriptions (in the form of metadata) of their content becomes crucial. Such descriptions will connect data publishers and consumers; publishers will advertise "what" there is inside a data source so that specialized applications can be created for data source discovering, cataloging, selection and so forth. Proposals like the VoID vocabulary touched this aspect. However, VoID mainly

[9] http://mapsforweb.wordpress.com/

focuses on providing *quantitative* information about a data source. We claim that toward comprehensive descriptions of data sources qualitative information is crucial. We introduce a theoretical framework for describing RDF data sources in terms of their completeness. We show how existing data sources can be described with completeness statements expressed in RDF. We then focus on the problem of the completeness of query answering over plain and RDFS data sources augmented with completeness statements. Finally, we present an extension of the completeness framework for federated data sources. The completeness reasoning framework is implemented in a tool available online[10].

3.21 Building Blocks for a Linked Data Ecosystem

Axel Polleres (Siemens AG – Wien, AT)

Linked Data has gained a lot of attention as a kind of "Silver Bullet" within the Semantic Web community over the past years. Still, some indications point at adoption not progressing as one might have expected optimistically around two years ago, when the standards an technologies around Linked Data seemed to be on the edge to "mainstream". In this short position statement argue for critical reflection on the state of Linked Data and discuss missing building blocks for an effective Linked Data Ecosystem.

The promise of Linked Data as a common data platform is the provision of a rather lightweight, standardised mechanism to publish and consume data online. By (i) using RDF as a universal, schema-less data format and (ii) "linking" to different datasets via re-using URIs global as identifiers,[11] Linked Data bears the potential of building a true "Web" of data. Apart from RDF [7], the accompanying RDFS&OWL [2, 6] standards enable the description and "linkage" of schema information in a loosely coupled manner, plus finally SPARQL [5] provides standard means to access and query Linked data. As such, Linked Data provides the basis for enabling Web "dataspaces" [9].

The so-called Linking Open Data (LOD) cloud diagram[12] documents the development of openly accessible Linked Data datasets between May 2007 and September 2011 (comprising 295 datasets). Now, almost two years after the last incarnation of the LOD cloud diagram has been published, the community behind Linked Data is faced with high expectations to proof added value and one might ask her/himself what has happened since then. While there has not been a new diagram published since September 2011, we may take the number of currently 339 LOD datasets[13] as an indication of developments since then, which may be viewed as at least a flattened growth rate in LOD. More worrying than the rate of growth (which may be hidden) though is actually the status of these LOD datasets. For instance, it seems that popular datasets that have been announced in 2010, such as the NYT dataset[14] have not been updated since then: the most recent value for the RDF

[10] http://rdfcorner.wordpress.com/
[11] 1 paraphrasing Linked Data principles, cf. http://www.w3.org/DesignIssues/LinkedData.html, retrieved June 2013.
[12] http://lod-cloud.net/, retrieved June 2013
[13] http://datahub.io/group/lodcloud
[14] http://data.nytimes.com/

property http://data.nytimes.com/elements/latest_use pointing to the latest article about an organization is "2010-06-14".

It is thus probably a good moment to take a step back to critically reflect on which puzzle pieces might be missing to achieve (even more) widespread adoption of Linked Data. Particularly, it seems that more than a few publishing principles and community enthusiasm is necessary to keep the idea of a Web-scale data ecosystem afloat. In the following, we will outline some challenges and missing building blocks to complement the available standards for Linked Data towards a fully functioning ecosystem.

Not all Linked Data is open. Particularly from an industry perspective, not all Linked Data will be open and available under open licenses: on the one hand, consumers will want to combine their own closed datasets with publicly available Linked Data; on the other hand, Linked Data may be published in different, non-compatible, possibly commercial licenses, which may impose restrictions on the use, re-use and re-publication of available data. In this context, it seems clear that the Linked Data community will need to provide standardised mechanisms to deal with access restrictions and different licenses, especially to make Linked Data interesting for industry. For starting points, cf. for instance Denny Vrandecic's recent writeup[15] or preliminary works on license composability via extending Semantic Web mechanisms [10].

Linked Data needs Mechanisms to deal with Dynamicity & Evolution. As the example in the introduction showed already, Linked Data may likely become outdated if not maintained properly. In this context, we note that there is a lack of standard technologies to both annotate temporal validity of evolving linked data as well as to process dynamic linked data as it evolves. We argue that standard technologies and best practices might help publishers to keep their data up-to-date and easy maintenance. Cf. for instance [8, 11].

Linked Data Quality needs Provenance & Trust. In order to determine trustworthiness and quality of Linked Data and combine data from different sources, it will be necessary to track provenance and trust, and to take these factors into account for query evaluation. Whereas the recent W3C PROV standard recommendation [4] provides a good starting point for describing and tracking provenance, integration with the remaining Linked Data standards and devising bespoke methods for query processing probably still needs more work.

Linked Data needs more (and less) than OWL. While not all features of OWL and particularly OWL2 seem to be equally adopted within published Linked Data [3], we note that a lot of published structured data is of numerical nature (e.g. public statistics). For integrating such data, different machinery than schema alignment supported via current ontology languages like RDFS and OWL is needed; rather, standard mechanisms for unit conversion or other mathematically expressible dependencies among properties are needed, cf. [1] for possible starting point.

The author is looking forward to discuss how these missing building blocks can be built and combined into a working ecosystem for Linked Data in the Dagstuhl seminar on "Interoperation in Complex Information Ecosystem". Starting points mentioned in the present position paper do not mean to be exhaustive and we shall be discussing further inputs. The author's expectation on the seminar is a road-map outlining

1. how these building blocks can implemented in terms of industry strength standards and best practices that can interplay and scale on the Web and
2. where further fundamental research is necessary.

[15] https://plus.google.com/104177144420404771615/posts/cvGay9eDSSK

References

1 Bischof, S., Polleres, A. *RDFS with attribute equations via SPARQL rewriting*. Proceedings of the 10th ESWC. vol. 7882, pp. 335–350. Montpellier, France, 2013.
2 Brickley, D., Guha, R. *RDF Vocabulary Description Language 1.0: RDF Schema*. W3C Recommendation, 2004. http://www.w3.org/TR/rdf-schema/
3 Glimm, B., Hogan, A., Krötzsch, M., Polleres, A. *OWL: Yet to arrive on the web of data?* Proceedings of WWW2012 Workshop on Linked Data on the Web (LDOW 2012), 2012.
4 Groth,P.,Moreau,L. *An overview of the PROV family of documents* W3C Recommendation, 2013. http://www.w3.org/TR/prov-overview/
5 Harris, S., Seaborne, A. *SPARQL 1.1 query language*. W3C Recommendation, 2013. http://www.w3.org/TR/sparql11-query/
6 Hitzler, P., Krötzsch, M., Parsia, B., Patel-Schneider, P.F., Rudolph, S. *OWL 2 Web Ontology Language Primer*. W3C Recommendation, 2009. http://www.w3.org/TR/owl2-primer/
7 Manola, F., Miller, E., McBride, B. *RDF Primer*. W3C Recommendation, 2004. http://www.w3.org/TR/rdf-primer/
8 Umbrich, J., Karnstedt, M., Hogan, A., Parreira, J.X. *Hybrid SPARQL queries: Fresh vs. fast results* Proceedings of ISWC 2012, pages 608–624, 2012.
9 Umbrich,J., Karnstedt,M., Parreira,J.X., Polleres,A., Hauswirth,M. *Linked Data and Live Querying for Enabling Support Platforms for Web Dataspaces*. Proceedings of the 3rd International Workshop on Data Engineering Meets the Semantic Web (DESWEB). Washington DC, USA, 2012.
10 Villata, S., Gandon, F. *Towards licenses compatibility and composition in the web of data*. Proceedings of ISWC2012 Posters & Demos, 2012.
11 Zimmermann, A., Lopes, N., Polleres, A., Straccia, U. *A general framework for representing, reasoning and querying with annotated semantic web data* JWS 12, 72–95, 2012.

3.22 Towards Future Web-based Aerospace Enterprises

René Schubotz (EADS Innovation Works – Munich, DE)

Aerospace industry is in the midst of a deep evolution of its industrial organization model. Refocusing on architect-integrator activities, product life-cycle modularization and outsourcing policies are major critical success factors and stimulate the emergence of virtual extended enterprises. This necessitates adequate integration and interoperation strategies between the architect-integrator and its risk sharing partners participating in the design and manufacturing of aerospace products.

Facing the abundance of stakeholders, schemata and systems in hitherto "non-information industry" enterprises, the Linked Data promise of provisioning a *single* common data platform is tantalizing and nurtures the long-term vision of consolidated and trustworthy (engineering) data spaces that integrate data in all phases of the product lifecycle, such as shape and geometry, various facets of its physical and functional description, structural and material properties, models of analyses, engineering trade-off decisions, manufacturing information, etc.

Yet, industrial uptake of Linked Data standards and technologies is timid. In the following, the author tries to indicate some impediments at different organizational levels.

- **On the inter-enterprise level**, pressing questions concerning data security, confidentiality, trust and provenance need to be addressed. Previous works [4, 3] investigate potential vectors of attack, however, an industrial-strength Web of Data requires substantially more hardening. Moreover, industrial businesses require a notion of Linked Closed Data [2] which is published with access and license restrictions and therefore demands standardizing access, authentication and payment protocols.
- **On the enterprise level**, the challenge is to integrate data and workflows of product lifecycle tools in support of end-to-end lifecycle processes. This requires the exploration of suitable integration techniques with minimalistic specification effort. First steps towards this direction have been taken in the form of a Linked Data Basic Profile [5], sparking considerable community interest [1].
- **On the specialty department level**, highly elaborate engineering design pipelines need to be captured. By exploiting the workflow paradigm for capturing the design of engineering workflows, and RDF to interlink the workflow, its specialty domain engineering tools as well as static and dynamic data sources, increased efficiency of design, engineering and evaluation activities becomes possible. To realize this goal, we need languages and execution environments [8] that enable scalable and flexible utilization and manipulation of computing and data resources.
- **On the level of the individual, working engineer**, user-centered data perspectives on the engineering data space should enable a wide range of interactions with respect to any engineering task. However, the working engineer is *not* a working ontologist. This necessitates adequate navigational languages in order to explore the data relevant for the task at hand, as well as intuitive user-interfaces [6, 7] making the consumption and publication of Linked Data light-weight, easy-to-use and easy-to-understand.

The author firmly believes in the applicability of the current web architecture in the extended industrial enterprise, and is looking forward to discuss and exchange views on the current trends, challenges, and state of the art solutions.

References

1 Open services for lifecycle collaboration. http://open-services.net/.
2 M. Cobden, J. Black, N. Gibbins, L. Carr, and N. Shadbolt. A research agenda for linked closed data. In *Second International Workshop on Consuming Linked Data (COLD2011)*, 2011.
3 A. Hasnain, M. Al-Bakri, L. Costabello, Z. Cong, I. Davis, T. Heath, et al. Spamming in linked data. In *Third International Workshop on Consuming Linked Data (COLD2012)*, 2012.
4 A. Hogan, A. Harth, and A. Polleres. Scalable authoritative owl reasoning for the web. *International Journal on Semantic Web and Information Systems (IJSWIS)*, 5(2):49–90, 2009.
5 A. J. Le Hors, M. Nally, and S. K. Speicher. Using read/write linked data for application integration–towards a linked data basic profile. *Fifth International Workshop on Linked Data on the Web (LDOW2012)*, 2012.
6 M. Luczak-Rösch and R. Heese. Linked data authoring for non-experts. In *Second International Workshop on Linked Data on the Web (LDOW2009)*, 2009.
7 R. Schubotz and A. Harth. Towards networked linked data-driven web3d applications. In *First International Workshop on Declarative 3D for the Web Architecture (Dec3D2012)*, 2012.

8 S. Stadtmüller, S. Speiser, A. Harth, and R. Studer. Data-fu: A Language and an Interpreter for Interaction with read/write Linked Data. In *Proceedings of the 22nd International Conference on World Wide Web*, pages 1225–1236, 2013.

3.23 LITEQ: Language Integrated Types, Extensions and Queries for RDF Graphs

Steffen Staab (Universität Koblenz-Landau, DE)

Joint work of Steffen Staab, Scheglmann Steffen, Gerd Gröner, Eveyne Viegas

Motivation

RDF data representations are flexible and extensible. Even the schema of a data source can be changed at any time by adding, modifying or removing classes and relationships between classes at any time. While this flexibility facilitates the design and publication of linked data on the Web, it is rather difficult to access and integrate RDF data in programming languages and environments because current programming paradigms expect programmers to know at least structure and content of the data source. Therefore, a programmer who targets the access of linked data from a host programming language must overcome several challenges. (i) Accessing an external data source requires knowledge about the structure of the data source and its vocabulary. As linked data sources may be extremely large and the data tend to change frequently, it is almost impossible for programmers to know the structures at the time before they develop their programs. Therefore, approaches to simplify access to RDF sources should include a mechanism for exploring and understanding the RDF data source. (ii) There is an impedance mismatch between the way classes (types) are used in programming languages compared to how classes structure linked data. (iii) A query and integration language must be readable and easily usable for an incremental exploration of data sources. (iv) When code in a host language describes how RDF data is to be processed by the resulting program, the RDF data should be typed and type safety should be ensured in order to avoid run time errors and exceptions. To address these challenges, we present LITEQ, a paradigm for querying RDF data, mapping it for use in a host language, and strongly typing it for taking full advantage of advanced compiler technology.

In particular, LITEQ comprises:

- The node path query language (NPQL), which has an intuitive syntax with operators for the navigation and exploration of RDF graphs. In particular, NPQL offers a variable free notation, which allows for incremental writing of queries and incremental exploration of the RDF data source by the programmer.
- An extensional semantics for NPQL, which clearly defines the retrieval of RDF resources and allows for their usage at development time and run time.
- An intensional semantics for NPQL, which clearly defines the retrieval of RDF schema information and allows for its usage in the programming environment and host programming language at development time, compile time and run time. Our integration of NPQL into the host language allows for static typing – using already available schema information from the RDF data source – making it unnecessary for the programmer to manually re-create type structures in the host language.

Discussion

LITEQ has been partially implemented as part of F#, full implementation is underway. LITEQ benefits from type providers in F# that support the inte- gration of information sources into F# [2] such that external data sources are directly available in programs. Type provider use F# LINQ queries [1] to retrieve schema information from (Web) data sources in order to build the corresponding types at run-time. Several Type Provider demonstrate the integration of large data sources on the Web, like the Freebase Type Provider that allows for the navigation within the graph-structure of Freebase[16]. The novel contribution of LITEQ is the full exploration of properties and the distinction of extensional and intensional use of the LITEQ query language. The core advantage of LITEQ compared to other integration approaches that map RDF into a host language is its integration of the different phases of exploration, compile and run time – benefitting both from the type definitions and the extensional queries.

Acknowledgements

This work has been supported by Microsoft.

References

1 Erik Meijer, Brian Beckman, and Gavin Bierman. *LINQ: Reconciling Object, Relations and XML in the .NET Framework*. In Proceedings of the Twenty-Fifth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, page 706, Chicago, Illinois, June 2006.

2 D. Syme, K. Battocchi, K. Takeda, D. Malayeri, J. Fisher, J. Hu, T. Liu, B. Mc-Namaa, D. Quirk, M. Taveggia, W. Chae, U. Matsveyeu, and T. Petricek. *F# 3.0 — Strongly Typed Language Support for Internet-Scale Information Sources*. Technical Report MSR-TR-2012-101, Microsoft Research, 2012.

3.24 Connecting Web APIs and Linked Data through Semantic, Functional Descriptions

Thomas Steiner (Google – Hamburg, DE)

Joint work of Verborgh, Ruben; Steiner, Thomas; Mannens, Erik; Van de Walle, Rik

In the beginning days of the Semantic Web, developers considered "services" as something separate from the "data" Web. There was a clear distinction between the weather offered by the webpage http://example.org/weather/saarbruecken/today and the weather offered through the endpoint found at http://example.org/services, which required you to construct a `POST` request with a body of method=getWeather&city=Saarbrücken. The former could be described as data, the latter had to be described as a service, while both were actually doing the exact same thing. Describing data could be done with simple RDF, whereas services needed to be described quite verbosely with OWL-S or WSMO.

Services built in the REST architectural style [3] do not exhibit this complexity, because the unit of a REST API is a *resource*, which is *identical* to the RDF concept of a resource. In

[16] The Freebase Wiki about the Schema: http://wiki.freebase.com/wiki/Schema

fact, "REST service" is a contradictio in terminis, because the purpose of a REST API is to *not* stand out as a service, but rather to expose application concepts as resources. It is a matter of hiding internal implementation details (which are subject to change anyway) in order to guarantee the evolvability of the exposed API. In a REST API, it doesn't matter if the weather data is generated by a service in the back or if it is simply a pre-generated document. These details are something only the server should care about. Therefore, exposing application concepts as resources is the responsible thing to do, and it should not come as a surprise that this works well in combination with RDF, which is after all an acronym for *Resource* Description Framework.

The Semantics of Change

With services and data both being resources, there's still a gap that needs to be bridged: what about state-changing operations? While data-providing services might be elegantly exposed as resources, data-modifying services might seem more difficult. However, that shouldn't be that case. The HTTP uniform interface foresees different methods for manipulation [1], including `PUT`, `POST`, `DELETE`, and recently also `PATCH`. While the semantics of almost all methods are strongly constrained, `POST` involves a degree of freedom, as "[t]he actual function performed by the `POST` method is determined by the server" [1]. This essentially means that the protocol does not allow one to predict what the result of the action will be. While this is not a problem for humans, who can interpret out-of-band information, it is a complex task for machines, who need to somehow *understand* what it means to perform a `POST` request on a certain resource. This made us wonder how we could describe the semantics of change in a machine-interpretable way.

To achieve this, we created the description format RESTdesc [5, 7], the goals of which are two-fold:

Capturing functionality RESTdesc descriptions capture the functionality of an HTTP request by connecting the pre-conditions and post-conditions of a given action in a functional way. They key to this connection are variables and quantification over these variables, functionality which is not supported natively by RDF. Therefore, RESTdesc descriptions are expressed in Notation3 (N3), a superset of RDF. The benefit here is twofold. First, the semantics are *integrated* into the language, as opposed to the use of expression strings in RDF, which are not supported natively. Second, when RESTdesc descriptions are instantiated, they become regular RDF triples, which can be handled as usual.

Describing the request Additionally, RESTdesc aims to explain the request that needs to be made to achieve the action, *without* harming the hypermedia constraint [2]. RESTdesc descriptions are merely a guidance, an expectation, but the interaction is fully driven by hypermedia, inheriting all the benefits of the REST architectural style (such as independent evolution).

Design for easy discovery and composition

The fact that RESTdesc descriptions are native N3 citizens makes them interpretable by any N3 reasoner. This means that any reasoner is able to solve the problem of discovery (which of the descriptions match a given need) and composition (combining different descriptions to match a need). Composition experiments conducted with RESTdesc and the EYE reasoner [4] show that even compositions with complex dependency chains can be created in a few hundred milliseconds. This number is largely unaffected by the total number of present descriptions.

RESTdesc adds the missing piece of the puzzle to have a seamless integration between data and services. While the REST principles make a universal treatment in the form of resources possible (all of which can be described by regular semantic technologies such as RDF), RESTdesc bridges the gap by describing the functionality of state-changing operations, which are an important aspect of Web APIs. Examples of RESTdesc usage can be found on http://restdesc.org/, together with an explanation of reasoner-based composition. A recent use case is distributed affordance [4], RESTdesc-based generation of hypermedia controls.

References

1 R. T. Fielding, J. Gettys, J. Mogul, H. Frystyk, L. Masinter, P. Leach, and T. Berners-Lee. Hypertext Transfer Protocol – HTTP/1.1. IETF Standards Track, June 1999.

2 Roy T. Fielding. REST APIs must be hypertext-driven. Untangled – Musings of Roy T. Fielding, October 2008.

3 Roy T. Fielding and Richard N. Taylor. Principled design of the modern Web architecture. *Transactions on Internet Technology*, 2(2):115–150, May 2002.

4 Ruben Verborgh, Vincent Haerinck, Thomas Steiner, Davy Van Deursen, Sofie Van Hoecke, Jos De Roo, Rik Van de Walle, and Joaquim Gabarró Vallés. Functional composition of sensor Web APIs. In *Proceedings of the 5th International Workshop on Semantic Sensor Networks*, November 2012.

5 Ruben Verborgh, Michael Hausenblas, Thomas Steiner, Erik Mannens, and Rik Van de Walle. Distributed affordance: An open-world assumption for hypermedia. In *Proceedings of the 4th International Workshop on RESTful Design*, May 2013.

6 Ruben Verborgh, Thomas Steiner, Davy Van Deursen, Sam Coppens, Joaquim Gabarró Vallés, and Rik Van de Walle. Functional descriptions as the bridge between hypermedia APIs and the Semantic Web. In *Proceedings of the 3rd International Workshop on RESTful Design*, pages 33–40. ACM, April 2012.

7 Ruben Verborgh, Thomas Steiner, Davy Van Deursen, Jos De Roo, Rik Van de Walle, and Joaquim Gabarró Vallés. Capturing the functionality of Web services with functional descriptions. *Multimedia Tools and Applications*, 64(2):365–387, May 2013.

3.25 Heterogeneous Information Integration and Mining

Raju Vatsavai (Oak Ridge National Laboratory, US)

With the entry of private satellite corporations, the number of satellites operating in sun synchronous orbits has increased in recent years. As a result, today we are dealing with big heterogeneous data that is multi-resolution, multi-spectral, multi- sensor, and multi-temporal in nature. Multitude of these heterogeneous data products allows us to overcome information gaps arising due to environmental conditions (e.g., clouds during inclement weather conditions) and multi-temporal imagery allows us to monitor both natural and man-made critical infrastructure. However, analyzing these big heterogeneous data products poses several challenges. First, there are no good statistical models for heterogeneous data that allows accurate classification and change detection. Existing models are primarily designed for similar attributes (e.g., Gaussian Mixture Models). Second, derived data products (e.g., land-use/land-cover maps) do not follow any standard classification scheme. Though some of these products can be integrated using ontologies (at attribute level), spatial union of these data products is still an open research problem. Specific research problems are listed below.

- Statistical classification/clustering models for heterogeneous data (e.g., optical and synthetic aperture data; or continuous random variables and discrete/multinomial attributes)
- Fusion of thematic maps: Ontology drive spatial integration (both attributes and spatial joins)
- Model fusion:
  - Distributed data sources: How to construct a global modal from local models (derived independently)?
  - Heterogeneous data: How to fuse models generated independently on each sensor product?
  - Multi-temporal data: How to fuse models generated on each temporal instance?
- Data fusion and reduction methods that preserve object boundaries (e.g., liner relationships in feature space)

Acknowledgements

Prepared by Oak Ridge National Laboratory, P.O. Box 2008, Oak Ridge, Tennessee 37831-6285, managed by UT-Battelle, LLC for the U. S. Department of Energy under contract no. DEAC05-00OR22725.

3.26 Semantics and Hypermedia Are Each Other's Solution

Ruben Verborgh (Ghent University – iMinds, BE)

Joint work of Verborgh, Ruben; Steiner, Thomas; Mannens, Erik; Van de Walle, Rik

The Linked Data principles versus the rest constraints

Linked Data, oftentimes referred to as *"the Semantic Web done right"*, starts from four simple principles, as stated by Tim-Berners Lee [1]:

1. Use URIs as names for things.
2. Use HTTP URIs so that people can look up those names.
3. When someone looks up a URI, provide useful information, using the standards.
4. Include links to other URIs, so that they can discover more things.

While these principles are starting to get known in the REST community as well, the main principles behind the REST architectural style [3] are mostly unknown to the semantic community. If you allow us to be creative with their phrasing and order, they are:

1. Any concept that might be the target of a hypertext reference has a resource identifier.
2. Use a generic interface (such as HTTP) for access and manipulation.
3. Resources are accessed through various representations, consisting of data and metadata.
4. Any hypermedia representation should contain controls that lead to next steps.

Since REST's flexible representation mechanism gives the freedom to publish information using standards, wouldn't you say that both sets of principles are actually pretty close? And considering the fact that the fourth Linked Data principle is the determining condition for the "Linked" adjective, and the fact that the corresponding fourth principle of REST indicates hypermedia controls as essential for REST APIs [2], doesn't it come to mind that both communities might actually be striving towards the same? Because it are exactly the *links* that give the data semantics and thus make it useful, and they're the same links that

drive REST APIs. After all, in a true REST API, you can perform all actions you need by following links, just like you do on the Web. This is called *"hypermedia as the engine of application state"*.

This insight leads us to conclude that the Linked Data principles are largely equivalent to the REST principles. In particular, we can see REST's necessity of links as an operational variant of the fourth Linked Data principle. For data, links are used to create meaning; for applications, links are used to perform actions. In both cases, they are essential to discover the whole world starting from a single piece.

The tragedy of the missing link

The difference between both approaches is the impact of missing links. Within Linked Data, the condition is that you should link your data to *some* URIs. If you forget to link to a certain set of (meta-)data, it doesn't matter all that much: the concept you are linking to might in turn link to the concept you forgot. Indeed, the linking concept is transitive, so the meaning a client is looking for can still be discovered. From the operational, REST point of view, things are different: if you are viewing a piece of information and you want to go to a certain place that is not linked, well... hypermedia gives up on you. It is then impossible to perform the desired action directly through the hypermedia document and you must do something else (like opening Google). This might only sound like a minor inconvenience, but it's more: why do we have hypermedia if we can't use its controls anyway, since they don't bring us to the place we want?

This is exactly the problem we are trying to tackle in our latest research. The omission of the links you need as a user is only natural, because how can the server possibly know what next steps you want to take? So if the server does not provide the *affordance* [2] to go to the place you need, the client must add it. In our architecture and implementation for *distributed affordance* [4], we automatically generate the hyperlinks the user needs in a personalized way. These links are constructed based on semantic annotations in the page *(thus, Linked Data)*, which are matched at runtime [5] to a user-selected set of services (REST). For instance, if you are reading a book review page, your browser can automatically generate links to borrow this book from your local library or to download it to your iPad. These links form personalized affordance for you, based on the content but selected according to your preferences.

So what happens here is that we connect two loose ends—the information on the one hand and the service on the other—solely based on semantics. This allows a loose coupling at design time (the information publisher does not need to know about the services you prefer) while having a strong coupling at runtime (the links directly let you use the information with the service). You could say that semantics complete the hypermedia engine: if the link is missing or does not exist, semantics can generate it. But let's be fair and also say it the other way: hypermedia completes semantics, by offering the services that allow you to do something you need with the data you like.

References

1 Christian Bizer, Tom Heath, and Tim Berners-Lee. Linked Data – The story so far. *International Journal On Semantic Web and Information Systems*, 5(3):1–22, 2009.

2 Roy T. Fielding. REST APIs must be hypertext-driven. Untangled – Musings of Roy T. Fielding, October 2008. http://roy.gbiv.com/untangled/2008/rest-apis-must-be-hypertext-driven.

3 Roy Thomas Fielding. *Architectural Styles and the Design of Network-based Software Architectures*. PhD thesis, University of California, Irvine, California, 2000.

4 Ruben Verborgh, Michael Hausenblas, Thomas Steiner, Erik Mannens, and Rik Van de Walle. Distributed affordance: An open-world assumption for hypermedia. In *Proceedings of the 4th International Workshop on* RESTful *Design*, May 2013. http://distributedaffordance.org/publications/ws-rest2013.pdf.

5 Ruben Verborgh, Thomas Steiner, Davy Van Deursen, Sam Coppens, Joaquim Gabarró Vallés, and Rik Van de Walle. Functional descriptions as the bridge between hypermedia APIs and the Semantic Web. In *Proceedings of the 3rd International Workshop on* RESTful *Design*, pages 33–40. ACM, April 2012. http://www.ws-rest.org/2012/proc/a5-9-verborgh.pdf.

4 Working Groups

4.1 Future of Actions

Réne Schubotz

Joint work of Valeria Fionda, Claudio Gutierrez, Armin Haller, Andreas Harth, Axel Polleres, Réne Schubotz, Ruben Verborgh

What if actions have their own URIs on the Web, are they getting executed every time you access the URI? Currently you can invoke actions via resource-centric web technologies and standards, however, you want to know beforehand what the resource is doing for you. Our assumption is that a standard solution for actions on the Web should include the following: Execution should not be possible by GET, you need to get a representation of the action, maybe get a redirect, have typed URIs (for results, execution) and content negotiation.

Our proposal is that we can use a standard protocol inspired by SPARQL endpoints, but to allow additionally to send the results to some other place, which is specified in the protocol and send with the query. This protocol can have a process model included that can delegate actions. Consequently, when you have a lot of data, you can define a URI where the data is stored, or some description of the data, rather than the direct request/response style that you have currently in the SPARQL protocol. As such resources are tasks, and the actions defined in the protocol can be part of a workflow. For example, if I order a flower, you post a job/action and a response URI and you can check the new resource that was created as part of your action to check if the flower was delivered to whomever you specified. We still need a process model for this protocol and the issue of notification is open. We also discussed that a publish/subscribe mechanism could potentially be implemented over this protocol.

4.2 Data Profiling

Felix Naumann

Joint work of Felix Bießmann, Christian Bizer, Kjetil Kjernsmo, Andrea Maurino, Felix Naumann

Data profiling is the process of extracting metadata from a given dataset. Application areas for data profiling include data mining, visualization, schema reengineering and query

formulation, query optimization, data integration and cleansing, data quality assessment, and data source discovery and description. In summary, profiling is an important preparation step for any other data-intensive task.

While traditional data profiling considers single relational databases or even only single tables, data profiling for data ecosystems broadens the profiling scope and tasks significantly. Data profiling can and should be applied at every level: Entire data spaces (many and diverse sources), individual data domains (many sources from single topic), data sources (many tables and classes), individual data classes (such as persons, places, etc.), and finally data properties (such as name, address, size, etc.).

The main challenges of data profiling for large data ecosystems are the definition and specification of which metadata to extract at which level, the aggregation of metadata across levels, and of course the actual computation of the metadata by analyzing the sometimes very large datasets.

4.3 Enterprise Semantic Web

Melanie Herschel

Joint work of Stefan Desßloch, Melanie Herschel, Bernhard Mitschang, Giovanni Tummarello

In a wide variety of applications, ranging from reporting to complex business analytics, enterprises traditionally rely on systems processing relational data efficiently and at scale. However, the highly structured nature of such data makes them difficult to link and integrate. As Semantic Web technologies, such as linked RDF data and means to efficiently query and process these become available, an interesting question is how these technologies may impact on Enterprise Data Management.

In studying the question, the first observation is that RDF data by itself is interesting in the Enterprise context, as such data bears information that may not be available in other sources. So, similarly to other types of data sources (XML, Excel, CSV), RDF data can be considered as one among many interesting data sources that will however be processed in a similar way as other types of data. Therefore, the remainder of the discussion focused on where in the process of integrating, manipulating, and analyzing data Semantic Web technologies may apply in the Enterprise context.

We identified two domains of particular interest to an Enterprise that may benefit from the Semantic Web, i.e., (1) Knowledge Representation for documentation or traceability purposes and (2) gaining Knowledge from (Big) Data. We decided to further elaborate on this second aspect, where the main advantage of Semantic Web is the fact that it may render the time-consuming design and maintenance of a global schema and complex schema mappings unnecessary. One avenue for future research is however to study the price of this simplification, susceptible to be paid at a different point in the process. For instance, relational query engines take advantage of the rigid structure of data to efficiently process large volumes of data, so we raise the question whether we can leverage the maturity of these systems while benefitting from the advantages of RDF data. More specifically, is there a hybrid data model taking the best of both worlds and can we design query-languages (and execution engines) that can seamlessly and efficiently deal with both types of data?

In moving from a highly structured type of data to (Linked) RDF data, we also observe the need to shift from the classical extract-transform-load paradigm used in data warehouses

to an extract-explore-analyze paradigm. Here, more mature tools for data profiling, browsing, visualization, etc. need to be developed before Enterprises switch to this paradigm. Also, Enterprises will only be willing to move to this new paradigm if they can be convinced of technological benefits in terms of business relevant key performance indicators such as return on investment or total cost of ownership. Therefore, we believe that scientific evaluation should take these into account in the Enterprise context.

4.4 Planning and Workflows

Ulf Leser

Joint work of Stefan Decker, Yolanda Gil, Melanie Herschel, Ulf Leser

The group discusses the impact of techniques from planning from first principles for scientific workflows. There was a general concern that such methods require a formal representation of properties of tasks, starting points and the targeted data product, which often are not available and also hard to maintain because tools and methods change very rapidly. The group then identified workflow repositories and workflow similarity search as another way to aid developers in designing workflows. Instead of generating workflows de-novo from abstract specifications, the idea here is to use similarity searches in workflow repositories to identify existing, proven workflows that solve the task at hand. If no perfect match is found, functionally similar workflows might help as starting point for adaptations. Workflow similarity search is a field with many existing and timely research questions, such as proper (semantic) similarity measures, methods for similarity-based workflow exploration and auto-completion, algorithms for searching across different workflow models, or workflow mining.

4.5 Linked Data Querying

Aidan Hogan

Joint work of Valeria Fionda, Katja Hose, Ulf Leser, Giuseppe Pirrò, Kai-Uwe Sattler, Steffen Staab

The ultimate destination of Linked Data is a decentralised eco-system of information spread over the Web. Following some Web standards (RDF, SPARQL, OWL, HTTP, URIs) and generic guidelines (dereference able URIs, provide links), publishers act independently when contributing to this information space. Aside from choosing a standard data-model, only loose co-ordination exists between publishers with respect to how data should be described, linked and made available. Although hundreds of Linked Datasets are now officially available on the Web – together comprising of billions of triples – it is still unclear what infrastructure is needed to query the data.

Current data-access mechanisms/problems include:

Dereferenceable documents: As per the definition of Linked Data, the URIs that name things should return useful information about those things when they are looked up over HTTP. Typically, all local URIs (i.e., URIs under the authority of the Linked Data site) should be dereferenceable and should return all RDF triples where the URI is in the

subject position. Some datasets also provide all RDF triples where the URI is in the object position. However, HTTP lookups are costly and typical queries can require large numbers of such lookups. Politeness policies (artificial delays) must be implemented by the client to avoid DoS attacks on the data provider. Furthermore, data that are not dereferenceable are difficult to access (short of a full site crawl).

Data Dumps: Many sites make data dumps available for download. Such dumps are often "all or nothing" in that the client can choose to download all of the data or none. Such data access mechanisms are of use to clients such as warehouses or analytical applications that wish to process/index the entire dataset locally. No standard protocols exist for accessing site dumps, other than sparsely-available VoID descriptions. Minor updates to data often require re-loading the full dump.

SPARQL Endpoints: Many Linked Data sites make SPARQL endpoints publicly available on the Web, which allows arbitrary clients to issue complex queries over RDF data hosted by the server. However, such endpoints suffer from performance and reliability problems. SPARQL is a complex query language (evaluation of which is PSpace-complete) and queries can be expensive to compute. Thus endpoints often time-out, return partial results, or fail under heavy loads.

During discussions at the Dagstuhl Seminar, we observed the following issues with respect to querying Linked Data:

- Protocols to find the overall "schema" of the data provided in a Linked Dataset are not available. Class and property URIs can be dereferenced individually, but the result is a collection of vocabulary definitions, not a structural overview of the data (cf. data profiling discussion).
- Creating structured queries is difficult since the client may not be familiar with the contents/vocabulary/structure of the dataset.
- Although Linked Data-providers adopt a common data structure, provide links and share some vocabulary definitions, this only enables a coarse form of client-side data integration: the integration problem is far from "solved".
- RDF stores are no longer so naive with respect to query optimisation. Various works using database optimisation techniques have been published in the (database) literature. Various benchmarks, competitions and commercial engines have also emerged. Some engines rely heavily on compression/in-memory techniques. Schemes for sharding/partitioning RDF datasets also exist, where indexes over billions, tens of billions and hundreds of billions of triples have been claimed. In summary: problems of querying Linked Data not solely attributable to naive/poorly designed RDF stores.
- Making SPARQL endpoints publicly available breaks new ground. In the database word, back-end SQL engines are rarely/never made open. Rather services and limited query interfaces (e.g., RESTful interfaces) are built on top.

Looking to the future of Linked Data Querying, we identified the following directions:

- SPARQL endpoints require excellent cost-model predictions to know a priori if they can service a query adequately or not, and to do the required load-balancing of requests from multitudinous/arbitrary Web clients. Without this, Quality-of-Service guarantees and load scheduling become very difficult.
- There is a clear trade-off between SPARQL endpoints, which push all of the processing on to the server, versus Dereferencable Documents/Site Dumps, which push all of the processing on to the client. There is nothing "in the middle". A simplified RDF query

language, perhaps expressed as a RESTful service, would perhaps strike a happy medium (working title: "GoldiLODs"). One idea was a query language that disabled join processing, only allowing atomic triple lookups and perhaps basic filtering and pagination mechanisms. This would allow the client to get a focused set (or sets) of data for local processing in a much more "controlled" fashion than traditional dereferencing. Furthermore, cost models, load balancing, hosting, etc. would be greatly simplified for the server, enabling a more reliable service than a SPARQL endpoint.

- Many expensive queries issued to SPARQL endpoints are analytical queries that require processing a large subset (or all) of the indexed data. It is not sustainable for the server to incur the costs of such queries. Languages for expressing analytical needs over RDF are still missing. One possibility is to create a procedural language designed for RDF analytics (perhaps allowing declarative SPARQL queries to be embedded).
- Guides to help clients create queries are also of importance. The RDFS and OWL vocabulary descriptions are only superficially helpful in the task of query generation. Other methods to explore the structure of the data on a high-level are needed to understand how queries are best formulated (i.e., with respect to which vocabulary elements to use, how joins are expressed, what sub-queries will lead to empty results, etc.).
- There is still no standard mechanism for full-text search over the textual content contained within RDF literals (SPARQL has REGEX but this is applied as a post-filtering operator, not a lookup operator).

4.6 Semantics of Data and Services

Sheila McIlraith

Joint work of John Domingue, Craig Knoblock, Sheila McIlraith, Giuseppe Pirrò, Raju Vatsavai

The objective of this breakout session was to address issues related to the semantics of data and services. Of particular concern was a means of defining the semantics of linked data. The primary motivation for addressing this topic is to facilitate the composability and interoperation of data with data, and data with services. We believe that some degree of semantic description for data and services is necessary to successful integration and interoperation of data and services. A second motivation for semantic descriptions of data was so that data sets could be suitably annotated, archived and found. This was viewed as being increasingly important for scientific work.

There has been substantial previous work on the topic of semantics for services in support of Semantic Web Services (SWS), including W3C proposals and recommendations such as OWL-S, WSMO, SWSO, SAWSDL, and most recently the Linked Data based WSMO-Lite, MicroWSMO, the Minimal Service Model and Linked-USDL based on SAP's Unified Service Description Language, as well as other efforts. The first three efforts, while differing slightly in their vocabulary and ontology formalism (OWL-S is in OWL, WSMO in WSML and SWSO is in first-order logic), each of these ontologies for services provides what we believe to be an adequate, or close to adequate, description of services. However, these early efforts results in complicated descriptions that are rarely created in full form and more recent work (e.g., iServe & Karma) has focused on simpler more streamlined source descriptions that are not as rich, but can be created automatically. One deficit we discussed was with respect to the description of the diverse forms of data that we may wish to integrate or interoperate

with. This includes a diversity of structured, semi-structured, unstructured and streaming data, and possibly the query engines that are employed to query that data. It was felt that further consideration needs to be given to a suitable description of such data.

There has also been significant previous work on tools to assist in the creation of specific service and data semantic annotations. These include SOWER for WSDL based services, SWEET for annotating Web APIs, and most recently Karma, which uses machine learning techniques to semi-automatically build source descriptions. There has also been significant previous work, too numerous to list, on the development of tools that exploit semantics of services and data. Some highlights include the SWS brokers WSMX and IRS-III, the Linked Data based semantic repository iServe (which is now part of the Linked Open Data Cloud), and the many OWL-S based tools. Finally, it was noted that Semantic Web Services has had broad influence in the development of tools by major corporations, including but not limited to IBM, SAP, and Apple through Siri.

Just as was the case with Semantic Web Services, the scope of the semantics is best circumscribed by what is needed to enable various applications. The tasks that informed the development of both OWL-S and WSMO included automated service discovery, invocation, composition, simulation, verification, mediation and execution monitoring. We felt these were also (at least) the tasks to be considered in any further effort.

Despite several SWS upper ontologies, few services and data contain annotations that describe their semantics. (That may be inaccurate/an overstatement.) As such, after a decade of SWS effort, it remains hard to find data and services. A number of service discovery tools have been developed in the past. The status of these tools needs to be explored, but it was predicted that they suffer from a lack of suitably annotated services and that the technology itself is still appropriate. We also felt there was a need for a data discovery engine. It was observed that it would be useful not only to search with respect to the topic of the data, but also with respect to its past use, and its provenance.

Challenges: The group identified the following challenges: 1. Defining a vocabulary/ontology for describing services/data. 2. Automatic methods for generating semantic descriptions of services. 3. How do you use Crowdsourcing/gamification to aid in the above. 4. How do we know if we're complete? Task dependent? 5. Calculating the domain and range of a service? 6. Need the relationship between the inputs and outputs. Easy to capture type information e.g. this service takes inputs of type A and then has outputs of type B. 7. Having descriptions which incorporate authentication is important as 80% of services require this e.g. on ProgrammableWeb. 8. Which representation language manage tradeoff between expressiveness/tractability. Existing Web standards. Do we need a KR at all? 9. Covering different types of data e.g. Linked Data and describing the semantics of streaming data. 10. How to link interpretations of data (e.g. satellite date) especially privacy preserving models. 11. Overcoming sampling problems if one can't access all the data.

4.7 Streams and REST

Daniela Nicklas

Joint work of Stefan Deßloch, Bernhard Mitschang, Daniela Nicklas, Kai-Uwe Sattler, Réne Schubotz, Thomas Steiner

How can streams of data, coming from active data sources, be integrated in the REST architecture pattern? There are many work-arounds for this problem, ranging from repeating polls to "negotiations" over REST interfaces and streaming over other protocols. We came up with an approach that implements the common pub/sub mechanism for continuous queries, uses the well-known REST term (although with slightly adapted semantics), works with any stream query language, and can even integrate prosumers (combined producers and consumers of data, e.g., a mobile application that sends its location updates and gets back continuous query results that depend on its location). We believe that this approach nicely fits to the REST paradigm, and future work would be to implement it in a running prototype.

4.8 Clean Slate Semantic Web

Giovanni Tummarello

Joint work of Aidan Hogan, Katja Hose, Steffen Staab, Giovanni Tummarello

Linked (Open) Data has been a movement launched circa around 2007 which put emphasis on the publishing of data instances on the web, stressing the importance of a universal, agreed data retrieval mechanism as an enabler for the overall "web of data" or "semantic web vision". While the objective of this initiative and the focus on a pragmatic connection between "data" and "existing web mechanism" is recognized as very positive, it is a fact that the initiative did not so far have much success, with the data growth apparently stopped, and well known issues in data availability, quality of the existing data – as well as lack of notable applications making use of it. The idea of the working group was to question the "Linked Data principles", proposed originally and to ask ourself what could be other principles that could lead to in higher incentive for quality data publication and reuse. The group did by no mean have the time to investigate the issues thoroughly but made observed that, with some respect, Linked Data principles (basically dereferenceable URIs) is broken:

For "simple lookups" – questions and answer cases:

- It's not useful for a client who doesn't know where to start. Before asking information about anything one should know the URI but no mechanism is provided to ask for it. E.g. how to enter a URI about "Dagstuhl"?
- It was never defined what exactly a response should be in terms of completeness of a response: sometimes due to data modelling "the triples attached to something" is a very insufficient representation.

For use cases that would require a complete knowledge of the datasets e.g. "what is the biggest city you know of" or "give me all your movies" – crawl all/warehouse cases:

- There is no mechanism for this e.g. no connection to a SPARQL endpoint (and SPARQL would itself be a bad idea given how easy it is to ask questions which are too intense on the server).
- Even worse, there is no support for the ability to at least crawl the content of the site with a guarantee that one could get the whole dataset and then be able to ask the above questions. There is no mandatory sitemap in Linked Data and even if there was there would be no mechanism to retrieve data about URIs which are not on the same domain name. E.g. if a web site had this triple in its database http://dbpedia/resource/Berlin isa city:NiceCity it is unclear how would anyone be able to get this triple using the "Linked Data principles".

We also questioned if the emphasis on URIs and the requirement for a "Linked Data" publisher to put links to other dataset is justified. Looking at reality would suggest that people – but most of all enterprises which are those who own the most important and often data rich web sites – would very seldom do anything that doesn't give them immediate benefits and in general on the web one should seek decoupling and a link is a strong coupling. An alternative to suggesting to link, and even suggesting that anything should have a stable URI, could be that of stressing the importance of a good, comprehensive entity description, in a way that to humans and machine linking algorithms alike it would be easy to make a connection – a connection which could be created dynamically and according to the specifications of a task at hand (e.g. certain tasks might consider certain kind of equality vs other tasks which would require other kind, e.g. looser or more strict.). We called this "pointing based referencing approach vs model-theory approach to referencing". We agreed however than when possible stable URIs are of course great and links are good for a client – given that for a lone client downloading a full site is out of the question.

We also mentioned that given big data and the power of search engines and web infrastructure one could easily see clients having some sort of back-end support to overcome these limitations, with great potential for useful functionalities (e.g. enter a website and immediately be suggested the most interesting pages on it by an external service who had previously crawled it all).

In the discussion, we hinted at use cases browsing, content management, data integration, enterprise data management, and stakeholders that might have a say in the determination of a new protocol to publish data (if found to be required): lay persons, scientists, developers, web site managers. For each of these one should consider the costs associated with any data publishing and data consumption methodologies and the rewards associated.

We discussed organizational issues and if it made sense to determine preferred centralized vocabulary but the consensus was these would emerge, or be de facto coordinated by important players, e.g. as in schema.org. We finally concluded with a very rough idea on how any new proposal should be evaluated: cost for stakeholders, fulfillment of use cases, compliance and natural match with existing web standards.

Participants

- Felix Bießmann
TU Berlin, DE
- Christian Bizer
Universität Mannheim, DE
- Stefan Decker
National University of Ireland – Galway, IE
- Stefan Dessloch
TU Kaiserslautern, DE
- John Domingue
The Open University – Milton Keynes, GB
- Valeria Fionda
Free Univ. of Bozen-Bolzano, IT
- Yolanda Gil
University of Southern California – Marina del Rey, US
- Claudio Gutierrez
University of Chile, CL
- Armin Haller
Australian National Univ., AU
- Andreas Harth
KIT – Karlsruhe Institute of Technology, DE
- Melanie Herschel
University of Paris South XI, FR
- Aidan Hogan
National University of Ireland – Galway, IE
- Katja Hose
Aalborg University, DK
- Martin Junghans
KIT – Karlsruhe Institute of Technology, DE
- Kjetil Kjernsmo
University of Oslo, NO
- Craig A. Knoblock
University of Southern California – Marina del Rey, US
- Ulf Leser
HU Berlin, DE
- Andrea Maurino
University of Milan-Bicocca, IT
- Sheila McIlraith
University of Toronto, CA
- Bernhard Mitschang
Universität Stuttgart, DE
- Felix Naumann
Hasso-Plattner-Institut – Potsdam, DE
- Daniela Nicklas
Universität Oldenburg, DE
- Giuseppe Pirrò
Free Univ. of Bozen-Bolzano, IT
- Axel Polleres
Siemens AG – Wien, AT
- Kai-Uwe Sattler
TU Ilmenau, DE
- Rene Schubotz
EADS – Ottobrunn, DE
- Steffen Staab
Universität Koblenz-Landau, DE
- Thomas Steiner
Google – Hamburg, DE
- Rudi Studer
KIT – Karlsruhe Institute of Technology, DE
- Giovanni Tummarello
National University of Ireland – Galway, IE
- Raju Vatsavai
Oak Ridge National Lab., US
- Ruben Verborgh
Ghent University, BE