Provenance and Formal Methods: The Case of Digital Image Processing

Carlos Sáenz-Adán
Departamento de Matemáticas y Computación, Universidad de La Rioja, Logroño, Spain

1 STAGE OF THE RESEARCH

There is a well know problem of balancing efficiency and reliability when researchers attempt to combine scientific computation with formal verification of algorithms. Usually, verified programs cannot compete in performance with respect to applications in production. Both aspects (efficiency and reliability) are in particular very important in bioinformatics applications (for instance, in the context of biomedical image processing).

With the aim of addressing these problems, this research aims to set up an environment in which both scientific computation with digital images and formal verification of their algorithms are combined using techniques and standards of provenance. This environment design could be able to facilitating the reproducibility of the processes, and also to explaining the reasons why that processing has been performed.

2 STATE OF THE ART

In the project entitled "Formalisation of Mathematics", included in the 7th European program called (ForMath), formal proof libraries were developed (using proof assistants such as Coq, Isabelle and ACL2) in several mathematical fields. One of the main aims of this project was the formalization of scientific computing algorithms for increasing trust in Symbolic Computation and Computer Algebra systems. In particular, the Spanish node of the project (led by the University of La Rioja team) made relevant contributions on homological processing of digital images (Lambán et al., 2014) (Poza et al., 2014) and PhD thesis (Poza, 2013)).

An important characteristic of the project is that a real collaboration with a biologists team, which studies synthesis of drugs for alleviating neurodegenerative diseases (such as Alzheimer), provides us with real examples for processing digital images. The company SpineUp (SpineUp), led by the researcher Miguel Morales, has posed problems from biomedical images that have been overcome by the Computer Science group at University of La Rioja. These contributions have been presented in the Spanish Neuroscience conference (Mata et al., 2013) (Mata et al., 2011). Some implemented algorithms used in these developments were formally verified within the ForMath project.

This research line which joins (scientific) computation and deduction (verification of programs and algorithms) has been recently granted by the Spanish Government (project MTM2014-54151-P).

In order to understand our interest in these issues, we must explain that the Computer Science group from University of La Rioja develops different lines of research. In this team, researchers from the fields of scientific computing and formal methods in Software Engineering coexist with others who come from research in Information Systems, and more specifically in the area of data and knowledge metamodeling, and with others coming from the processes and workflow management fields.

Modeling can be related to provenance. This relationship comes from the concept of *Occurrence-Oriented system*, which has been fruitfully exploited in collaboration with the Noesis research group at the University of Zaragoza, led by Eladio Domínguez (Domínguez et al., 2014). Regarding process management, previous contributions are related to service oriented architectures (SOA), specifically devoted to ensure an agreed security policy covering the whole chain, from the service establishment to its consumption (Rodriguez-Priego and García-Izquierdo, 2007).

All these interests converge to the provenance field. This is a huge research topic, with fast growth and not mature enough at this moment. In principle, several ways have been considered to organize provenance. On the one hand, the provenance of information is considered (data-oriented workflows). It has been studied in the case of formal theorem proving (Ikeda et al., 2013) (which is related to the ForMath project), in the case of execution of programs (Cheney et al., 2011) (also using formal methods (Acar et al., 2013)), and finally in the case of databases (Cheney et al., 2009). Data provenance and workflow provenance may be also distinguished (Buneman and

Sáenz-Adán C..
Provenance and Formal Methods: The Case of Digital Image Processing.

Davidson, 2010). As discussed below, in our case both data provenance (digital images) and the processes involved in dealing with data (image processing) are relevant.

3 RESEARCH PROBLEM

The research line applying formal methods in image processing is being very successful but have some drawbacks (well known in the international community working in these topics). This weakness is related to how combining efficiency (required in real applications, in particular in bioinformatics) with reliability (increased by verifying programs using proof assistants). Reliability is an important property in scientific computing (in particular, in the area of biomedicine), but it conflicts with the pursuit of efficiency. It is well known that verified programs cannot compete in speed with applications in production at scientific laboratories (see, for example, (Poza et al., 2014)).

In this research we want to reduce the distance between deduction and final applications. To this aim, we propose to use information system techniques, in particular modern techniques coming from provenance.

4 EXPECTED OUTCOME

Our proposal, which to the best of our knowledge is new in the literature, would consist of including in a single network of provenance both causal chains (in other words, those that are producing the various transformations in digital images) and chains of arguments (those that explain why a certain process has been implemented). This would not only facilitate the reproducibility of experiments (one of the explicit aims of provenance), but also reproduce the reasons why different workflow steps have been designed, making it easier for an external agent the understanding of the process as a whole.

On the more basic case, a conceptual relationship of provenance could be a reference to an external document (for example, a technical report, a well-known algorithm or a journal paper) that explains why the associated functional process is consistent (with respect to the workflow objectives). In a more formalized context, the reference may contain a formal proof on a theorem prover (Coq, Isabelle, ACL2, etc.) showing that the algorithm is correct. Finally, in addition to a formal proof, a verified program could be available as an “explanatory artifact”. This certified program (whose correctness is ensured by means of a theorem prover) could be applied to obtain results that can be compared against the outputs obtained in the computing part of the provenance net.

In a paradigmatic case, an agent could consult the network of provenance to know what programs are applied for a certain image (including the query of previous images used as parameters, if any), may request the re-calculation (with the selected software platform) and in addition (and this is the most innovative aspect of our proposal) one could make an automated testing against a verified program that could show that the production program is appropriate in that particular case. Since the verified program can be (and, in general, will be) more inefficient than the programs in production, one can consider performing off-line testing, with the aim of not harming the agility of the reproduction of the functional part of the experiment.

It should be noted that our proposal avoids the problem (intractable with the current state of the technology) of formal verification of algorithms of scientific computing in production, since we do not look for ensuring that the operational program is equivalent, for all inputs, to the verified program. However, it allows us an flexible and uncoupled integration of formal methods that would increase trust in the experiment as a whole.

5 OUTLINE OF OBJECTIVES

Based on the problems identified above, we have defined the following four objectives:

- Objetive A. Definition of a setting where deduction and processing provenance coexist.
- Objetive B. Devising of a representation language which integrates (scientific) computing and deduction (verified algorithms and programs).
- Objetive C. Set up a query language over the previous representation.
- Objetive D. Develop a proof of concept. We are going to use the particular case of homological processing of images, with a prototype which is going to be useful to provide a provenance net. This is going to include both already-made formal proofs and ex-novo proofs.

6 METHODOLOGY

The proposed project is multidisciplinary, hence, we will use, from a methodological point of view, tech-

niques and methods from different scientific fields. It will be necessary to use the most appropriated one in each stage of development.

For instance, the basic techniques of literature search can be combined with more advanced techniques as the systematic review (Kitchenham et al., 2002). When tasks related to mathematics are addressed, we have to use more formal methods. That may be combined with methods from mechanized theorem proving. Since the project will likely require software development and systems integration, it will also be necessary to apply methods from the design of information systems and software engineering, such as requirements analysis and conceptual modeling.

6.1 Stages

Each stage corresponds with one objective, and has been split in several sub-stages.

- E1. Define a contextual environment with the aim of integrating deduction and computing provenance.
- E1.1. Study previous works in the research group, related both to formal verification of algorithms and to information systems.
- E1.2. Systematic review of the literature related to provenance.
- E1.3. Study of the expressiveness of different proposals in the literature, trying to adapt some of them (or a mixture of several ones) to achieve our objectives.
- E1.4. Set up a semi-formal definition of a provenance model which allows integrating the workflow of a process from a functional perspective, together with explanations describing why the process has been produced in that way.
- E2. Set up a formal definition of a language which represents models corresponding to the previous stage.
- E2.1. Study of different languages inside the literature to represent provenance networks (at least PLM (Del Rio et al., 2010), OPM (Moreau et al., 2011) and W3C Prov (Missier et al., 2013)).
- E2.2. Propose a representation language with the aim of dealing with the second objective.
- E3. Define a query and definition language for networks constructed with the previous representation language.
- E3.1. Analyze the available tools (in particular developed by the group, such as RCM (Rodriguez-Priego et al., 2013)) for managing data and processes.
- E3.2. Formal definition of a query language for provenance networks.
- E4. Development of a prototype which will use the proposals and definitions mentioned above.
- E4.1. Deployment, in a particular network, of some of the processes already developed for the manipulation of biomedical images, including formal proofs with Isabelle / HOL, Coq or ACL2.
- E4.2. Development of new features with formal proofs.
- E4.3 Justifying that the prototype can also integrate new sources of data and arguments developed in the previous stage.

6.1.1 Schedule

Based on the objectives, a doctoral planning has been done. It has been divided into four stages, each one corresponding to one year.

Throughout the PhD planning there are in addition tasks on coordination and supervision meetings with thesis advisors and other members of the research group. It is also foreseen the participation in training courses and conferences to expose partial results obtained.

Furthermore, there will be tasks related to documentation generation (internal reports, journal and proceeding papers) and to the development of programs and formal poofs.

REFERENCES

Acar, U. A., Ahmed, A., Cheney, J., and Perera, R. (2013). A core calculus for provenance. *Journal of Computer Security*, 21(6):919–969.

Buneman, P. and Davidson, S. B. (2010). Data provenance–the foundation of data quality. *)ˆ(Eds.):‘Book Data provenance–the foundation of data quality’(2013, edn.)*.

Cheney, J., Ahmed, A., and Acar, U. A. (2011). Provenance as dependency analysis. *Mathematical Structures in Computer Science*, 21(06):1301–1337.

Cheney, J., Chiticariu, L., and Tan, W.-C. (2009). *Provenance in databases: Why, how, and where*, volume 4. Now Publishers Inc.

Del Rio, N., da Silva, P. P., and Porras, H. (2010). Browsing proof markup language provenance: Enhancing the experience. In *Provenance and Annotation of Data and Processes*, pages 274–276. Springer.

Domínguez, E., Pérez, B., Rubio, Á. L., Zapata, M. A., Lavilla, J., and Allué, A. (2014). Occurrence-oriented design strategy for developing business process monitoring systems. *Knowledge and Data Engineering, IEEE Transactions on*, 26(7):1749–1762.

ForMath. http://wiki.portal.chalmers.se/cse/pmwiki.php/ForMath/ForMath.

Ikeda, R., Das Sarma, A., and Widom, J. (2013). Logical provenance in data-oriented workflows? In *Data Engineering (ICDE), 2013 IEEE 29th International Conference on*, pages 877–888. IEEE.

Kitchenham, B. A., Pfleeger, S. L., Pickard, L. M., Jones, P. W., Hoaglin, D. C., El Emam, K., and Rosenberg, J. (2002). Preliminary guidelines for empirical research in software engineering. *Software Engineering, IEEE Transactions on*, 28(8):721–734.

Lambán, L., Rubio, J., Martín-Mateos, F.-J., and Ruiz-Reina, J.-L. (2014). Verifying the bridge between simplicial topology and algebra: the eilenberg–zilber algorithm. *Logic Journal of IGPL*, 22(1):39–65.

Mata, G., Cuesto, G., Morales, M., Rubio, J., and Heras, J. (2011). Synapcountj: un software para el estudio de la densidad sináptica. In *XIV Congreso de la Sociedad Española de Neurociencia (SENC 2011)*. http://www.senc2011.com/docs/programa_senc2011.pdf.

Mata, G., Fernández, P., Romero, A., Rubio, J., Cuesto, G., and Morales, M. (2013). Nucleusj: desarrollo de un plugin en fiji para el análisis de modelos de muerte neuronal. In *XV Congreso de la Sociedad Española de Neurociencia (SENC 201)*. http://www.senc2013.com/.

Missier, P., Belhajjame, K., and Cheney, J. (2013). The w3c prov family of specifications for modelling provenance metadata. In *Proceedings of the 16th International Conference on Extending Database Technology*, pages 773–776. ACM.

Moreau, L., Clifford, B., Freire, J., Futrelle, J., Gil, Y., Groth, P., Kwasnikowska, N., Miles, S., Missier, P., Myers, J., et al. (2011). The open provenance model core specification (v1. 1). *Future Generation Computer Systems*, 27(6):743–756.

Poza, M. (2013). *Certifying homological algorithms to study biomedical images*. PhD thesis, Universidad de La Rioja.

Poza, M., Domínguez, C., Heras, J., and Rubio, J. (2014). A certified reduction strategy for homological image processing. *ACM Transactions on Computational Logic (TOCL)*, 15(3):23.

Rodriguez-Priego, E. and García-Izquierdo, F. J. (2007). Securing code in services oriented architecture. In *Web Engineering*, pages 550–555. Springer.

Rodriguez-Priego, E., García-Izquierdo, F. J., and Rubio, Á. L. (2013). References-enriched concept map: a tool for collecting and comparing disparate definitions appearing in multiple references. *Journal of Information Science*, page 0165551513487848.

SpineUp. http://spineup.es.