Toward a Cloud-Based Integration of IR Tools

Allan Hanbury and Mihai Lupu

Institute of Software Technology and Interactive Systems
Vienna University of Technology
Favoritenstraße 9–11/188
A-1040 Vienna, Austria
hanbury@ifs.tuwien.ac.at, lupu@ifs.tuwien.ac.at

Abstract

This position paper presents the case for creating a cloud-based infrastructure for IR. The infrastructure should provide access to multiple components of IR systems and a method to easily integrate them in various configurations, as well as data and parameters used in IR experiments. Workflow tools are promising for flexible integration of components and sharing of configurations, and have already been adopted in multiple areas of computational science. This infrastructure should lead to better reproducibility of IR experiments, easier take-up of research results by industry and more effective IR evaluation.

1 Introduction

A very large number of software components for Information Retrieval are currently available, ranging from comprehensive toolboxes for indexing and searching (e.g. Solr/Lucene, Lemur/Indri, Terrier) to tools for specific tasks, such as lemmatizers, stemmers and part-of-speech taggers. It is possible to combine these components in multiple ways in the creation of a search engine, but this in general involves much work. First the software must be downloaded, compiled and brought to a functional state before the integration of components into a new constellation can take place. This means that a significant amount of time is wasted on "non-research" tasks before the actual research can begin [ACMS12]. The research part then often leads to modifications of the programs, modifications which are not always made available to the research community. This creation of locally-implemented IR systems results in poor reproducibility of published experimental results by other research groups, as the exact experimental system is generally difficult to reproduce elsewhere.

In the computational sciences in general, little focus has been directed toward the reproducibility of experimental results, raising questions about their reliability [FS12]. There is currently work underway to counter this situation, ranging from presenting the case for open computer programs [IHGC12], through creating infrastructures to allow reproducible computational research [FS12] to considerations about the legal licensing and copyright frameworks for computational research [Sto09].

A promising way to make the IR components available in a standard way is to provide them on a cloud infrastructure to be called through an API. Search engines could then be relatively rapidly created by researchers wishing to experiment with various combinations of components. Reproducibility could be guaranteed by distributing a specification of the component setup used in the experiments in a publication, and ensuring that the data and all experimental parameters (including queries, relevance judgements, etc.) are available on the

In: M. Salampasis, N. Fuhr, A. Hanbury, M. Lupu, B. Larsen, H. Strindberg (eds.): Proceedings of the Integrating IR technologies for Professional Search Workshop, Moscow, Russia, 24-March-2013, published at http://ceur-ws.org

cloud. The use of standardised datasets made available on the cloud would enhance the ease of reproducibility of experiments.

Workflows have already gained acceptance in other branches of science as a method for distributing the setup of a computational process, and should also be considered for adoption in IR. The use of workflows is discussed briefly in Section 2. The availability of these multiple IR services would also revolutionise IR evaluation, as discussed in Section 3. Finally, an advantage of making IR components more accessible is the possibility for them to be easily taken up and experimented with by companies potentially wishing to use them in search systems. Section 4 presents considerations on the cloud infrastrutures to use for all aspects of the proposed approach.

2 Workflows

Scientific workflows have recently emerged as a paradigm for representing and managing complex distributed scientific computations. They orchestrate the dataflow across the individual data transformation and analysis steps, as well as the mechanisms to execute them in a distributed environment [GDE+07, GdR09]. Workflows have the potential to function as a unifying framework for the integration of IR tools.

Initial steps to using workflows in IR and annotation were taken by Corubolo et al. [CWH07], who developed a framework using the Kepler workflow tool combined with the Chesire3 search engine. This work is however no longer accessible, and was done at an early stage in the development of scientific workflow tools, when more primitive tools were available.

In order to advance the use of workflows in IR, it will be necessary to create components for the workflow implementing commonly used approaches in the IR pipeline (preferably based on open source software, although executable components without source code are also conceivable). The creation of workflows for IR applications has the following advantages:

- Connection with existing IR tools
- Sharing of workflows leading to better reproducibility of experiments
- Facilitation of component-based evaluation
- Rapid prototype development

A number of open source workflow tools are available [DGST09]. A promising candidate is Taverna[1], as it is the most widely used tool on the myExperiment portal for sharing scientific workflows[2]. Taverna also has the advantage that it has been integrated with the U-Compare UIMA-based text mining and natural language processing system[3] [KDN+10].

With workflow techniques, it will be possible to set up IR experiments consisting of multiple standard components rapidly. When investigating domain-specific search problems, the flexibility will be available to adapt the workflow to the domain based on domain-specific data and knowledge. The rapid prototyping capabilities provided by the proposed infrastructure will also be useful in rapid design and testing of user interfaces, in particular those interfaces designed specificially for users performing highly-specialised search tasks.

3 Evaluation

While evaluation is well-established in IR research, it suffers from a number of drawbacks in the way it is currently implemented [Rob08]. The drawbacks include only evaluating full systems [HM10], poor reproducibility of experiments and poor choice of baselines for comparison of the results of experiments [AMWZ09]. The availability of multiple executable components, potentially linked into a workflow system, along with data and experimental parameters stored in the cloud, would revolutionise IR evaluation [ABB+12]. At a basic level, it would simplify research on the effect of constituent components of an IR system on the performance of the full system [KE11]. Furthermore, it would be possible to relatively easily recreate an IR system published in a paper, and then begin further research from this baseline. It is even conceivable that an automated evaluation using multiple configurations of components could run continuously, providing a flow of experimental results on the

[1] http://taverna.org.uk
[2] http://www.myexperiment.org
[3] http://u-compare.org

performance of many component configurations analogous to how large experiments in physics (e.g. the Large Hadron Collider) continuously produce data.

However, evaluation will also have to be done from another point of view. For a researcher or developer implementing an IR system for a specific set of data in a specific domain, it would be useful to be able to evaluate which of the many components of a certain type available are optimal for that specific task. For example, given a dataset to index, a description of the expected queries and results, which of the available stemmers would provide the best performance? Being able to answer this type of question requires evaluation guidelines going beyond the approaches currently standard in evaluation campaigns.

4 Cloud-based Experimentation

The cloud has innovated a number of aspects of computing, as it provides the appearance of infinite computing resources available on demand, eliminates up-front commitment by cloud users and provides the ability to pay for the use of computing resources on a short-term basis as needed [AFG+10]. The abilities necessary for the approach described in this paper are:

- Provide the ability to centrally store and make available large datasets — Cloud providers already provide this service. For example, Amazon hosts public datasets free of charge[4].
- Allow multiple users to process the stored data without requiring the data to be transfered elsewhere — this is done through linking virtual storage drives to computing instances as required. For example, Amazon public datasets are accessed in this way.
- Allow users to share executable components implemented in computing instances with other users — An approach for doing this will have to be developed.

There are a number of challenges in developing a suitable approach for the final point above. One can imagine that snapshots of computing instances containing executable components are made available in a type of "app store" for re-use. For open source software, the snapshots should also allow access to the code. In order to encourage the use of the components in research, they should be made available free-of-charge for this purpose (i.e., researchers would have to pay for computing time, but no license fees). However, if a company decides to adopt a component in a system used commercially, then researchers should receive compensation of some kind.

A further important consideration is who should provide this service. Commercial cloud providers are already able to provide it, but choosing a single commercial provider could result in a "lock-in" of IR research to a single provider, due to incompatibilities between services provided by different companies. Potentially, a publicly-funded cloud infrastructure would be good for running IR research and evaluation experiments. However, this would make the take-up by industry more complex as commercial services could not be run on this infrastructure. A possible solution is an interface from the publicly-funded infrastructure to multiple commercial cloud infrastructures allowing researchers to transfer components to these infrastructures for take-up by industry.

5 Conclusion

There is currently a push to make the results of computational science more reproducible. If IR wishes to remain at the forefront of this development, then it is necessary to implement an experimental infrastructure for IR. Making executable components for IR systems as well as data and experimental parameters available on a cloud-based infrastructure, along with tools such as a workflow infrastructure, is a good step toward fully reproducible IR experiments. A first step toward carrying out evaluation on the cloud is being taken in the recently started VISCERAL project [HML+12]. In VISCERAL, the focus is on taking advantage of the cloud to allow evaluation to be done on multiple Terabytes of data, avoiding the necessity of first downloading the data (i.e. bringing the algorithms to the data instead of the data to the algorithms). However, the creation of the experimental IR infrastructure discussed in this paper has the ability to revolutionise IR experimentation beyond what is being considered in the VISCERAL project.

Acknowledgements

The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreements 318068 (VISCERAL) and 258191 (PROMISE).

[4] `http://aws.amazon.com/publicdatasets/`

References

[ABB+12] M. Agosti, R. Berendsen, T. Bogers, M. Braschler, P. Buitelaar, K. Choukri, G. M. Di Nunzio, N. Ferro, P. Forner, A. Hanbury, K. Friberg Heppin, P. Hansen, A. Järvelin, B. Larsen, M. Lupu, I. Masiero, H. Müller, S. Peruzzo, V. Petras, F. Piroi, M. de Rijke, G. Santucci, G. Silvello, and E. Toms. *PROMISE Retreat Report: Prospects and Opportunities for Information Access Evaluation*. PROMISE Network of Excellence, 2012. http://www.promise-noe.eu/promise-retreat-report-2012/.

[ACMS12] James Allan, Bruce Croft, Alistair Moffat, and Mark Sanderson. Frontiers, challenges, and opportunities for information retrieval: Report from SWIRL 2012. *ACM SIGIR Forum*, 46(1):2–32, 2012.

[AFG+10] Michael Armbrust, Armando Fox, Rean Griffith, Anthony D. Joseph, Randy Katz, Andy Konwinski, Gunho Lee, David Patterson, Ariel Rabkin, Ion Stoica, and Matei Zaharia. A view of cloud computing. *Commun. ACM*, 53(4):50–58, 2010.

[AMWZ09] Timothy G. Armstrong, Alistair Moffat, William Webber, and Justin Zobel. Improvements that don't add up: ad-hoc retrieval results since 1998. In *CIKM '09: Proceeding of the 18th ACM conference on Information and knowledge management*, pages 601–610. ACM, 2009.

[CWH07] Fabio Corubolo, Paul Watry, and John Harrison. Location and format independent distributed annotations for collaborative research. In *Research and Advanced Technology for Digital Libraries*, volume 4675, pages 495–498. Springer Berlin / Heidelberg, 2007.

[DGST09] Ewa Deelman, Dennis Gannon, Matthew Shields, and Ian Taylor. Workflows and e-science: An overview of workflow system features and capabilities. *Future Generation Computer Systems*, 25(5):528–540, 2009.

[FS12] Juliana Freire and Claudio T. Silva. Making computations and publications reproducible with VisTrails. *Computing in Science & Engineering*, 14(4):18 –25, August 2012.

[GDE+07] Y. Gil, E. Deelman, M. Ellisman, T. Fahringer, G. Fox, D. Gannon, C. Goble, M. Livny, L. Moreau, and J. Myers. Examining the challenges of scientific workflows. *Computer*, 40(12):24–32, dec. 2007.

[GdR09] C. Goble and D. de Roure. The impact of workflow tools on data-centric research. In Tony Hey, Stewart Tansley, and Kristin Tolle, editors, *The fourth paradigm: data-intensive scientific discovery*, pages 137–145. Microsoft Research, 2009.

[HM10] Allan Hanbury and Henning Müller. Automated component-level evaluation: Present and future. In *Multilingual and Multimodal Information Access Evaluation*, volume 6360 of *Lecture Notes in Computer Science*, pages 124–135. Springer, 2010.

[HML+12] Allan Hanbury, Henning Müller, Georg Langs, Marc Weber, Bjoern Menze, and Tomas Fernandez. Bringing the algorithms to the data: Cloud-based benchmarking for medical image analysis. In *Information Access Evaluation. Multilinguality, Multimodality, and Visual Analytics*, volume 7488 of *LNCS*, pages 24–29. Springer, 2012.

[IHGC12] Darrel C. Ince, Leslie Hatton, and John Graham-Cumming. The case for open computer programs. *Nature*, 482(7386):485–488, February 2012.

[KDN+10] Yoshinobu Kano, Paul Dobson, Mio Nakanishi, Jun'ichi Tsujii, and Sophia Ananiadou. Text mining meets workflow: linking u-compare with taverna. *Bioinformatics*, 26(19):2486–2487, 2010.

[KE11] Jens Kürsten and Maximilian Eibl. A large-scale system evaluation on component-level. In *Advances in Information Retrieval*, volume 6611 of *Lecture Notes in Computer Science*, pages 679–682. Springer Berlin / Heidelberg, 2011.

[Rob08] Stephen Robertson. On the history of evaluation in IR. *Journal of Information Science*, 34(4):439–456, 2008.

[Sto09] V. Stodden. The legal framework for reproducible scientific research: Licensing and copyright. *Computing in Science & Engineering*, 11(1):35 –40, February 2009.