A Cloud-based Approach for Evaluation on Private Data

Allan Hanbury* and Henning Müller†

Abstract

This paper points out a number of drawbacks of the Strawman proposal on the workshop website. It then describes the cloud-based approach that will be used for evaluation in the VISCERAL project, and finally discusses some potential ideas that could be taken over from the VISCERAL approach for evaluation on private data.

1 Introduction

Much useful and potentially life-saving information could be gained through processing large groups of medical records. One of the most famous cases is discovering the increased probability of coronary heart disease associated with the Vioxx medication, through an analysis of patient records from Kaiser Permenente [2]. Feufel et al. [1] even go so far as to say that it is "in some cases unethical" to store such data without installing access mechanisms. However, privacy is an important concern in this further use of medical record information. Various ways of anonymising or pseudonymising medical record information have been developed [7], but there are still extensive concerns on the part of the medical record owners about information leakage. This paper considers the possibility of doing information retrieval (IR) evaluation without the necessity of giving the developers of the IR systems access to the private data.

This paper initially points out a number of potential disadvantages of the strawman proposal for an IR evaluation task on medical records given on the workshop website[1] (Section 2). It then presents an approach to evaluation on private radiology data that will be adopted in the VISCERAL[2] project [4] (Section 3). Finally, some potential influences of the VISCERAL approach on the proposed task are outlined (Section 4).

2 Potential drawbacks of the strawman proposal

This section discusses some of the potential disadvantages of the strawman proposal on the workshop website. The main disadvantage is that every participant in such an evaluation campaign would have to take on many of the tasks that are traditionally done by the evaluation campaign organiser, thereby increasing the threshold for participation. This means that the advantage of participants of benchmarks in that well-curated data sets can be obtained with little effort is not valid anymore. These tasks are:

- obtaining access to data (via a request to the ethics committee) and reformatting it;
- query development (for this case, most likely recruit experts to do this) that should better rely on large scale data and not on single persons;

*Institute of Software Technology and Interactive Systems, Vienna University of Technology, Austria, hanbury@ifs.tuwien.ac.at

†University of Applied Sciences Western Switzerland, henning.mueller@hevs.ch

[1] http://sigir12pdc.wordpress.com/strawman/

[2] http://www.visceral.eu online starting from late 2012

- create pools to have a limited set of documents to judge per topic;
- relevance judgements (most likely experts required for this too), which can be expensive;
- install and run one open source search engine per participant on the data (this task is made more arduous here due to the requirement to install each participant search engine locally — in standard evaluation campaigns this step is not necessary as the participants only use their own search engines); some of the open source systems might be working well on Linux and others on Windows making it difficult to run all of them on the same platform.

As pointed out in [5], "Test collections are very expensive to build, with the relevance judgments being the most costly." Many of the traditional synergy advantages due to participation in evaluation campaigns are lost with what is in effect parallel organisation of multiple evaluation campaigns. While this might be acceptable for data such as e-mail that participants own in abundance and are in a position to create sensible queries and judge relevance relatively easily, the non-negligible tasks of putting together a test collections and in addition organising and potentially paying experts to do the query generation and relevance judgements may discourage participation.

Furthermore, the requirement that the participants must make their search engines open source will likely discourage companies from participating. This could also not be seen as a problem, as one can learn little that is scientifically useful from the results of a proprietary search engine. However, companies working in this field not participating could also be seen as a disadvantage, as part of the spectrum of possibilities is not covered.

3 VISCERAL approach

VISCERAL is an EU-funded project that will begin in November 2012. It has the task of organising two rounds of evaluation on radiology 3D image data. The first round will be focussed on locating specific structures in the radiology images, while the second round is on retrieval of similar cases for diagnosis aid, based on visual and text (radiology report) modalities. Mainly due to the huge amounts of data to be used (datasets in the order of 10TB), it was decided to adopt a cloud-based approach, as this allows the participants to bring their code to the data stored on the cloud, instead of distributing the data to the participants to be stored and processed locally.

The approach that will be adopted for the retrieval for diagnosis aid round is now described. A small "training" dataset, consisting of anonymised radiology images and associated anonymised radiology reports will be made available to participants on a cloud service. The radiology image data is anonymised by removing identifying information and blurring faces, while the radiology report text is anonymised by removing all names of people and organisations[3]. Participants will each have instances of virtual machines running on the cloud service on which they can install their software (including a choice of proprietary libraries) and connect to the training dataset for adapting the indexing and retrieval software to the dataset, as illustrated in Figure 1. Both linux and windows instances will be available so participants can select their operating system of choice. Once source code has been compiled then the code could be removed if the developers do not wish to share it. Particularly in image analysis the use of a large number of existing libraries that are often open source is frequent. Tools such as MatLab, Octave and itk (Insight toolkit) can make system development much easier and faster. On the other hand this use of many libraries would make it hard to have the system installed in an environment where versions of compilers and basic libraries cannot be controlled easily.

On the deadline for submission, the instances in which the participant software is installed will be transferred to the organisers and participants will lose their access to these instances. The organisers will then link the instances to a large "test" dataset that was not available to the

[3] Radiology reports are considered relatively easy to anonymise as they are basically a report by a radiologist based on viewing a set of images, but usually without having read much additional information on the patient beforehand.

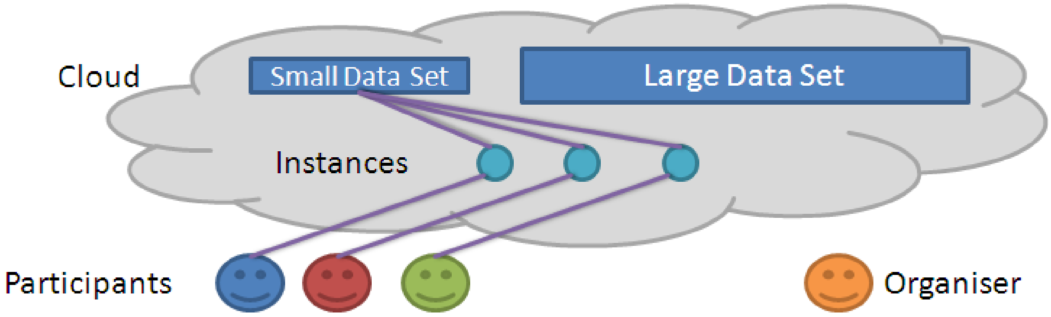

Figure 1: Training phase.

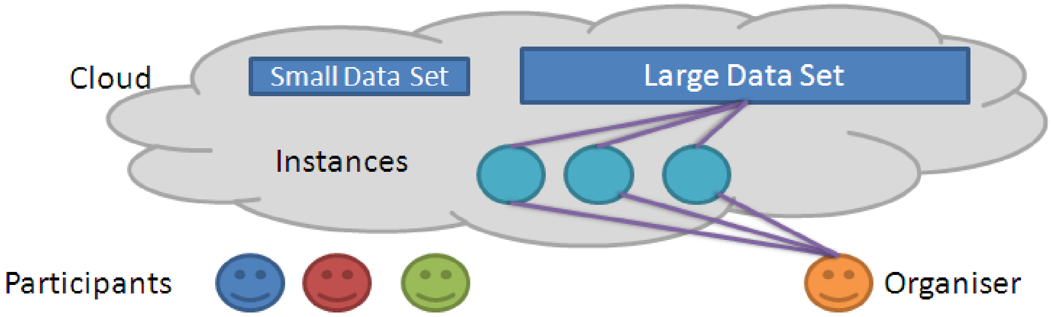

Figure 2: Test phase.

participants, as illustrated in Figure 2. The organisers will run the indexing software, followed by queries that have been generated on the dataset by radiology experts. Pooling will be done with relevance judgements by the experts, which will allow retrieval performance of each of the participant systems to be evaluated. For this evaluation campaign, the "large" dataset will also be anonymised as the data will be stored on a commercial cloud infrastructure.

4 Potential influences

The approach used in VISCERAL has the potential to overcome many of the drawbacks listed in Section 2. In the VISCERAL approach, the test dataset is never seen nor accessed by participants, so it could be non-anonymised private data. If it is private data, then the approach of using a commercial cloud provider is not suitable — it is likely that a private cloud solution with enhanced security features would have to be deployed inside the premises of the data provider. However, with this approach the standard evaluation campaign setup with one dataset, one set of queries, pooling over multiple systems for relevance judgements, etc. is feasible. Such an approach would also allow participants to participate without having to make source code available. Because participants need to get their code to work themselves inside the instances, the burden on the organisers of getting submitted code to compile and execute is removed, as is the requirement for participants to get the systems of other participants to function.

A drawback with this approach is that sufficient training data must be made available to participants. Exactly how this training data should be generated and the minimum amount necessary is a matter for discussion. A potential solution is to use the anonymised medical record data from the previous TREC medical track, reformatted into the format of the test data. Technical solutions could also be considered, such as giving participants the ability to run their systems on the private data and get evaluation metrics results back in a way that does not reveal any information from the data — for example the system could create the index in a partition that is not accessible to the participants. However, such a black box approach to the data is likely to be cumbersome for the participants, who will not be able to examine retrieved results during development. An alternative is to encourage the participants to obtain full access to medical record data for the development and training (as in the current strawman proposal), while evaluation is done on the hidden test set.

Participants who are willing to share components with others could also work together making tools available via standard interfaces as all participants work on the same infrastructure. This can be independent of whether source code is made available or not as web services could be employed for this. Such a component sharing can also help with component-based evaluation [3, 6] and make system development more efficient for participants.

5 Acknowledgements

The ideas for these evaluation campaigns arose from work that has received funding from the European Union Seventh Framework Programme (FP7/2007–2013) under grant agreement no. 257528 (KHRESMOI).

References

[1] M. A. Feufel, G. Antes, J. Steurer, G. Gigerenzer, J. A. Muir Gray, M. Mäkelä, A. G. Mulley, Jr., D. E. Nelson, J. Schulkin, H. Schünemann, J. E. Wennberg, and C. Wild. What is needed for better health care: Better systems, better patients or both? In G. Gigerenzer and J. A. Muir Gray, editors, *Better Doctors, Better Patients, Better Decisions: Envisioning Health Care 2020*, pages 117–134. MIT Press, 2011.

[2] David J Graham, David Campen, Rita Hui, Michele Spence, Craig Cheetham, Gerald Levy, Stanford Shoor, and Wayne A Ray. Risk of acute myocardial infarction and sudden cardiac death in patients treated with cyclo-oxygenase 2 selective and non-selective non-steroidal anti-inflammatory drugs: nested case-control study. *The Lancet*, 365(9458):475–481, 2005.

[3] Allan Hanbury and Henning Müller. Automated component-level evaluation: Present and future. In *Proceedings of the Conference on Multilingual and Multimodal Information Access Evaluation*, volume 6360 of *LNCS*, pages 124–135. Springer, 2010.

[4] Allan Hanbury, Henning Müller, Georg Langs, Marc-André Weber, Bjoern H. Menze, and Tomàs Salas Fernandez. Bringing the algorithms to the data: Cloud-based benchmarking for medical image analysis. In *Proc. of the CLEF Conference*, LNCS. Springer, 2012.

[5] Donna Harman. Information retrieval evaluation. *Synthesis Lectures on Information Concepts, Retrieval, and Services*, 3(2):1–119, 2011.

[6] 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.

[7] Thomas Neubauer and Johannes Heurix. A methodology for the pseudonymization of medical data. *International Journal of Medical Informatics*, 80(3):190–204, 2011.