Benchmarking: A catalyst to advance the state of the art in multimedia search computing

Martha Larson, PetaMedia Network of Excellence
Delft University of Technology
m.a.larson@tudelft.nl
Henning Müller, Chorus+ Coordination Action
University of Applied Sciences Western Switzerland (HES-SO)
Henning.mueller@hevs.ch

Benchmarks are valued for their ability to streamline research by eliminating redundancy, enabling direct performance comparison between algorithms, increasing efficiency by sharing resources between research sites and providing a concrete framework in which researchers interact in a productive mixture of competition and collaboration. This contribution discusses the benefits of benchmarking and explains the value of benchmarking for advancing the state of the art in multimedia search computing.

What is a benchmark?

A benchmark is a forum that organizes tasks for the research community. Researchers are invited to develop algorithms that address the tasks. In general (according to the so-called Cranfield paradigm [5]), a benchmarking task consists of three parts: 1. A task definition that describes the problem to be solved 2. A data set provided to the benchmark participants 3. Ground truth against which participants' algorithms are evaluated. Often benchmarks follow a yearly cycle that moves through a series of steps: identifying critical research challenges, developing specific tasks to address these challenges, releasing data, receiving results submitted by participants, evaluating and reporting results, convening participants in a workshop to discuss results and start planning the next year.

What are the benefits of benchmarking?

The benefits of benchmarking fall into two categories: benefits for the research community [3,6] and economic benefits gained by bringing innovative research closer to market [2]. Benchmarking benefits the research community by reducing fragmentation of research agendas and by promoting efficient use of resources. Without benchmarking initiatives, researchers develop solutions working isolated in their labs. Alone, they must identify innovative, high impact problems on which to concentrate their efforts and suffer from a lack of access to the larger, overarching perspective. They must create their own data sets from the ground up in order to evaluate their results. Individual datasets developed at individual research sites involve not only expending duplicate effort, but, more importantly, blocks the possibility of cross-site comparison. Comparison and reproducibility are two factors that drive forward the state of the art: if researchers know exactly where they stand with respect to the state of the art they can better direct their efforts to surpass it and more quickly abandon less promising lines of investigation. Benchmarking creates momentum: A researcher can turn to the benchmarking community for support in overcoming practical roadblocks, which are otherwise potentially both discouraging and time consuming. Finally, the mixture of competition and collaboration within a benchmarking initiative provides motivation for the group and makes the payoff of the invested effort immediately and clearly visible, both with and beyond the research community. Sharing data sets not only means that results are better comparable but by combining efforts much larger and thus statistically more relevant data sets can be created.

The economic benefits of benchmarks lie in their ability to coordinate efforts more closely between the research community and the needs of businesses bringing products to markets. In the initial phase of the benchmarking cycle where tasks for the upcoming year are planned, industry can contribute ideas for new tasks. Further, industry can also make datasets available that will allow researchers to focus on real world versions of specific tasks. In this way, algorithms developed within the benchmark are already closely linked to real-world business needs and are better suited for industrialization. In some industries such competitions are publicly announced challenges with a price to win such as a quality improvement in vide recommendation created in [9].
In order to build capacity in business, it necessary not only to innovate solutions, but also to formulate new problems. A solution to a new problem can provide the basis of a new product with the power to open up new market sectors. The flexibility and efficiency of benchmarking initiatives makes them uniquely suited to quickly address innovative new tasks and to come to a well-supported conclusion about the feasibility of addressing those tasks backed up by a large research community that is participating. Working together with a benchmark, a company is able to explore a new area that would normally be too risky or too expensive for it to tackle itself in addition to the difficulty of having state-of-the-art knowledge in the domain within the company.

Recently, explicit effort has been devoted to measuring the benefits of benchmarking. This effort has been led by large benchmarks sponsored by the US National Institute of Standards and Technology (NIST)[1]. In particular, the scholarly impact of the TRECVid[2] Video Retrieval Evaluation has been demonstrated by a bibliographic study [3] and a similar study has now been started for the ImageCLEF bechmark [6]. Further, an extensive report is now available on the economic impact of Text REtrieval Conference (TREC) Program [2]. An analysis of the impact of benchmarking data sets also in the long run can be found in [8]. Such results surely also extend to benchmarking initiatives that are organized from within Europe.
In remainder of this section, we turn specifically to the European perspective and discuss the importance of benchmarking initiatives for Europe.

Benchmarking to reinforce Europe's competitiveness

Europe's ability to compete internationally in ICT is dependent on the success of research initiatives that cross country boundaries. The MediaEval benchmark [1] is a prime example of an initiative that is able to leverage and focus research carried out at individual sites in individual countries, helping to ensure the efficient use of research investment. The MediaEval initiative is sponsored by the PetaMedia Network of Excellence[3] and the European Institute for Innovation and Technology's ICT Labs[4], two

[1] http://www.nist.gov/
[2] http://trecvid.nist.gov/
[3] http://www.petamedia.eu/

entities dedicated to fostering innovation and excellence and for bringing research results closer to market. MediaEval is an ongoing initiative and current information can be found on the website[5]. The MediaEval initiative allows for flexible participation – research sites can easily join the initiative and participate in tasks as dictated by their research needs and current capacity. The inclusive and dynamic organization of MediaEval is designed to help dissolve impediments to spontaneous cross-border cooperation and allow Europe to more effectively act as a single, unified research community.

The MediaEval initiative makes it possible to promote research topics that are critical for Europe within the international research community. It is clearly important for European researchers to take part in international research initiatives whose organization is based abroad. However, in order for Europe to maintain its defining role in international research agendas, it is necessary not only to participate in benchmarks, but also be actively involved in planning and organizing them. One particular aspect is the existence of ICT research areas that are of more central importance in Europe than they are in Asian or America. Europe can retain its key position by continuing to be the central international initiator of benchmarking efforts in theses areas. An example is the importance of developing technologies that make it possible to access multimedia content resources across cultures and independently of language. Here, Europe's own Cross-Language Evaluation Forum (CLEF[6]) has shown the value of bundling research effort in the area of text and image retrieval. Multilinguality and cultural differences are important in a European context and Europe is clearly a leader in these fields of information retrieval

MediaEval was originally established within the CLEF campaign and subsequently split off in order to become a stand-alone benchmark focused on multimedia in social contexts. Speech and language issues remain central for MediaEval. Unlike other multimedia initiatives, such as TRECVid, MediaEval is decoupled from investment from governments beyond Europe and in particular from US defense and intelligence spending. Although international cooperation is clearly important, in order to preserve the diversity of the goals pursued within the international research community, it is important to ensure the existence of independent initiatives that are fully free to focus on research areas central to creating more prosperous lives for the citizens and Europe and increased capacity for its businesses.

Similarly to MediaEval, ImageCLEF[7] [7] started within the CLEF benchmark but has more focused on various aspects of multimedia retrieval and visual information analysis. It has remained part of CLEF despite a focus more on visual media and language independentness than on several languages, although multilinguality is present in several sub tasks. ImageCLEF has mainly been a voluntary effort of many people that have changed over time. Since 2010 a partial support is given by the Chorus+ and Promise projects. ImageCLEF has run over ten different tasks since its start in 2003 and has raised the number of registered participants from 4 in 2003 to over 110 in 2011.

[4] http://eit.ictlabs.eu/

[5] http://www.mediaeval.org/

[6] http://www.clef-campaign.org/

[7] http://www.imageclef.org/

Automatic annotation of images as well as retrieval combining visual and textual means including cross-language retrieval have been the focus.

Challenges in benchmarking

Despite the clear advances of many science domains through benchmarks there are also critical voices stating that benchmarks can block innovation. When too much importance is put onto pure performance and not on novel technologies it can be better to make slight modifications to existing techniques and not develop something very novel, so this critic is clearly valid to a certain degree. Thus, it is important to stress the collaboration rather than the pure performance and remove importance from the goal of winning. This also means that benchmarks should continuously change and modify the tasks, so focus is put onto real innovation. MediaEval follows this concept of radically novel tasks such as the current focus on social media. This on the other hand can limit participation with research groups having to adapt to the challenges and the tasks. ImageCLEF has major changes for each task at least every three years to avoid being technologically locked but keeping a large participation. The impact analysis [3] also shows that tasks become cited less usually after three or four years, loosing their impact. Organizing a task in connection with a larger conference also has challenges as there are several constraints to take into account when organizing tasks across communities.

Several tasks are still challenges for benchmarks at the moment, for example the way towards extremely large datasets that may be hard to distribute to participants. It also means that data analysis becomes increasingly sparse and to find clear trends and compare techniques becomes even harder.
Another critical point is the yearly circle of benchmarks in phases. It gives researchers on the one hand a clear frame to focus on, but on the other hand it also blocks researchers to test new tools and techniques and their impact instantly. Continuous evaluation would allow to test automatically or semi automatically at any time, and could allow to adapt systems much quicker and more effectively.
Component-based evaluation is another important topic for the future as currently mainly entire systems are tested, whereas every retrieval system contains a large number of existing components such as image analysis or text retrieval tools. Measuring the overall performance can hide many important results and allowing to measure performance of components can help understanding the interactions and implications of these components.

Conclusions

In conclusion, this section has discussed the benefits of benchmarking both for the research community and for industry. Not only are these benefits potentially high, they are also high-value since the efficiency of benchmarking means that they can be achieved at relatively low costs. European efforts in the organization of benchmarks are necessary in order to secure the Europe’s international leadership in its key research area. The MediaEval and ImageCLEF benchmarking initiative make a contribution to this goal and give forum for the research community as well as to industry.

References

[1] M. Larson, M. Soleymani, P. Serdyukov, S. Rudinac, C. Wartena, V. Murdock, G. Friedland, R. Ordelman, and G.J.F. Jones. Automatic Tagging and Geotagging in Video Collections and Communities. ACM International Conference on Multimedia Retrieval (ICMR 2011), Trento, Italy, 2011.

[2] National Institute of Standards and Technology, 2010. Economic Impact Assessment of NIST's Text REtrieval Conference (TREC) Program. Final Report, July 2010. http://trec.nist.gov/pubs/2010.economic.impact.pdf

[3] F. Smeaton, P. Over, and W. Kraaij, 2006. Evaluation campaigns and TRECVid. In MIR '06: Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval, pages 321-330, 2006.

[4] C. V. Thornley, A. C. Johnson, A. F. Smeaton, H. Lee. The scholarly impact of TRECVid (2003–2009). J. Am. Soc. Inf. Sci. 62 (4), 613-627, 2011.

[5] C. W. Cleverdon, Report on the Testing and Analysis of an Investigation into the Comparative Efficiency of Indexing Systems, Cranfield College of Aeronautics, Cranfield, England, 1962.

[6] T. Tsikrika, A. Garcia Seco de Herrera, H. Müller, Assessing the Scholarly Impact of ImageCLEF, Cross Language Evaluation Forum (CLEF 2011), Amsterdam, The Netherlands, 2011.

[7] H. Müller, P. Clough, T. Deselaers, B. Caputo, ImageCLEF – experimental evaluation in visual information retrieval, Springer, 2010.

[8] M. Sanderson. Test Collection Based Evaluation of Information Retrieval Systems. Foundations and Trends in Information Retrieval, Vol. 4, No. 4, (2010) 247–375

[9] The Netflix Prize on video recommendation. http://www.netflixprize.com/, 2006.