Report on the SIGIR 2013 Workshop on Benchmarking Adaptive Retrieval and Recommender Systems

Pablo Castells
Univ. Autónoma de Madrid
28049 Madrid, ES
pablo.castells@uam.es

Frank Hopfgartner
TU Berlin
10587 Berlin, DE
frank.hopfgartner@tu-berlin.de

Alan Said
CWI
1098 XG Amsterdam, NL
alan@cwi.nl

Mounia Lalmas
Yahoo! Labs Barcelona
08018 Barcelona, ES
mounia@acm.org

1 August 2013

Abstract

In recent years, immense progress has been made in the development of recommendation, retrieval, and personalisation techniques. The evaluation of these systems is still based on traditional information retrieval and statistics metrics, e.g., precision, recall and/or RMSE, often not taking the use-case and situation of the actual system into consideration. However, the rapid evolution of recommender and adaptive IR systems in both their goals and their application domains foster the need for new evaluation methodologies and environments. In the Workshop on Benchmarking Adaptive Retrieval and Recommender Systems, we aimed to provide a platform for discussions on novel evaluation and benchmarking approaches.

1 Introduction

As a core topic in the foundation of IR and confluent areas such as Recommender Systems, evaluation is addressed by a significant stream of research on a yearly basis, making up a non-negligible share of the technical program of SIGIR, related front line conferences, and campaigns such as TREC, NTCIR, CLEF, etc. The evaluation of adaptive retrieval systems has been acknowledged to find difficulty in fitting to well-established evaluation paradigms and methodologies (e.g., [1]), which can be identified as a hurdle to research and progress in this area. Active research efforts and open discussion are currently taking place in parallel in the recommender systems and adaptive information retrieval fields, where devising methodologies and metrics suiting the goals and task models of real applications is still a prominent open issue.

The SIGIR Workshop on Benchmarking Adaptive Retrieval and Recommender Systems (BARS)[1] was intended to be a platform to revise and leverage the latest advances in this area, to identify the main issues to be addressed, and to share ideas for continued progress. The main motivation of the workshop was to join forces and provide a meeting point for researchers working on largely overlapping and connected areas such as adaptive retrieval and recommender systems, dealing with closely related problems but often from different backgrounds.

The workshop took place on 1st August 2013 as a half-day workshop in conjunction with ACM SIGIR'13 in Dublin, Ireland. The schedule included one keynote presentation, followed by two full paper presentations, and a concluding panel discussion on benchmarking adaptive retrieval and recommender systems. The workshop proceedings will be available on CEUR-WS. In the remainder of this report, we provide further details about the different parts of the workshop and end with a brief discussion on the future of the topic.

2 Keynote

In order to get insights into the evaluation of recommender algorithms in industry, we invited Torben Brodt, Head of Data Engineering at plista GmbH[2] to give a keynote presentation. plista is a data-driven content and ad distribution company located in Berlin, Germany. Their service recommends millions of news articles and ads every day to even more users. In his talk entitled *"The Search for the Best Live Recommender System"*, he presented challenges and opportunities that arise from handling vast amounts of user interaction data in real time for recommending news articles. He illustrated that context such as news categories, users' current geo-location or the day of the week play an important role in providing succesful recommender algorithms. Adressing the main focus of the workshop on benchmarking such algorithms, he talked about the News Recommender Challenge [4] which is organised as part of ACM RecSys 2013. Within this challenge, researchers can evaluate their recommendation algorithms in real time and using real user feedback[3]. In the context of this challenge, the organisers released a dataset consisting of 84 million news articles in real time. In [2], they outline the dataset, arguing that it can be used by the participants to benchmark and fine-tune their news recommendation algorithms. Another overview is provided by Said et al. [3]. Moreover, he explained that as part of the mentioned challenge, participants are able to see over a period of four weeks how popular their recommendations are with respect to users' clicking behaviour.

3 Full Paper presentations

The succeeding session was dedicated to the presentation of research papers on the subject matter. Presentations took twenty minutes, each followed by a ten minute discussion. Two papers got accepted for presentation at BARS'13:

Nadia A. Najjar from University of North Carolina at Charlotte presented her paper *"Tradeoffs in Evaluation Strategies for Group Recommender Systems"* which she co-authored

[1] `http://bars-workshop.org/`

[2] `http://plista.com/`

[3] Note that the CLEF-NEWSREEL lab of CLEF 2014 is a follow-up of this challenge. For more details, the reader is referred to `http://clef-newsreel.org/`.

with David C. Wilson from the same institute. She presented a survey of different types of evaluation of recommender systems for groups and focused on benchmarking the performance of different approaches.

In the second paper, Saurabh Gupta from IIT Madras presented his paper (co-authored with Sutanu Chakraborti) on *"Evaluating Conversational Recommender Systems based on Preference Based Feedback"*. He proposed to enhance the evaluation of simulation-based conversational recommender systems on preference-based feedback. The enhancements include a) an elaboration on the user behaviour model to include a probabilistic component, beyond just a similarity-driven behaviour, and b) incorporating the quality of recommendation (in addition to the time to reach a relevant target) during interactive recommendation sessions.

4 Panel Session

In order to discuss benchmarking of adaptive retrieval and recommender systems further, we invited various experts to join our panel session. We were happy to welcome Paul N. Bennett (Microsoft Research), Neil Hurley (University College Dublin), and Jimmy Lin (University of Maryland) who agreed to share their expertise with the workshop attendees. All panelists were able to highlight aspects from slightly different points of view with Jimmy Lin focusing on his experience while working at Twitter, Neil Hurley representing the point of view of the recommender systems community, and Paul Bennett representing the IR community.

The panelists stressed the importance of user interface and subjective aspects in the effectiveness of user adaptations. Presentation biases can overshadow internal system features and adaptive algorithms; beyond the system-based evaluation viewpoint, the effectiveness of personalised features is to a lagre extent about how users feel when they use the system. Explanations should not necessarily describe what the system did, but rather convey explanations that users can relate to. Everyone agreed on the general difficulty of evaluating adaptive IR systems and recommenders. Adaptive techniques involve several stakeholders, in scenarios where different performance measures are at play which may even conflict. The tension between tractability and reflection of real business and user goals was also pointed out. While quantitative measures have traditionally dominated in this area, the user experience perspective needs to be further brought into the effectiveness assessments, but the evaluation approaches need to be kept still simple and tractable. A/B testing is the de-facto method in industry, checking for signals such as click-through rate (follow-through in Twitter), and community engagement. Post-hoc activity monitoring beyond this is a key part of evaluation in this perspective, in order to determine how much subsequent activity can be attributed to a personalised recommendation (such as a who to follow suggestion on Twitter). The need for an observation period is often challenging in this context.

5 Conclusion

This workshop was a first attempt to join forces and provide a meeting point for researchers working on largely overlapping and connected areas such as adaptive information retrieval and recommender systems, dealing with closely related problems but often from different backgrounds. Concluding from the active engagement of the workshop attendees, we argue that the workshop was a good forum to trigger discussions on the evaluation of both adaptive IR and recommender systems. Being a workshop organised in conjunction with the annual

SIGIR conference, the majority of participants are from the information retrieval community. In the future, we intend to organise a second instalment of this workshop in conjunction with a more recommender systems dominated conference to further raise awareness of the similar research challenges that the two research communities are facing.

Acknowledgements

We would like to thank the organisers of SIGIR for providing a venue for this workshop. Further, we acknowledge the efforts of the members of the programme committee and additional reviewers, including: Dyaa Albakour (University of Glasgow, UK), Leif Azzopardi (University of Glasgow, UK), Alejandro Bellogín (CWI, NL), Nicholas Belkin (Rutgers University, USA), Toine Bogers (RSLIS, DK), Pia Borlund (RSLIS, DK), Paolo Cremonesi (Politecnico di Milano, IT), Noriko Kando (National Institute of Informatics, JP), Alexandros Karatzoglou (Telefónica, ES), Benjamin Kille (TU Berlin, DE), Bart Knijnenburg (UC Irvine, USA), Udo Kruschwitz (University of Essex, UK), Nikos Manouselis (Agroknow, GR), Martha Larson (TU Delft, NL), Neal Lathia (University of Cambridge, UK), Denis Parra (University of Pittsburgh, USA), Ian Ruthven (University of Strathclyde, UK), Domonkos Tikk (Gravity R&D, HU), Ryen White (Microsoft, USA), and Michelle Zhou (IBM, USA).

References

[1] Nicholas J. Belkin. Some(what) grand challenges for information retrieval. *SIGIR Forum*, 42(1):47–54, 2008.

[2] Benjamin Kille, Frank Hopfgartner, Torben Brodt, and Tobias Heintz. The plista dataset. In *NRS'13: Proceedings of the International Workshop and Challenge on News Recommender Systems*, pages 14–21. ACM, 10 2013.

[3] Alan Said, Jimmy Lin, Alejandro Bellogín, and Arjen de Vries. A month in the life of a production news recommender system. In *Proceedings of the CIKM Workshop on Living Labs for Information Retrieval Evaluation*, LivingLab. ACM, 11 2013. to appear.

[4] Mozhgan Tavakolifard, Jon Atle Gulla, Kevin C. Almeroth, Frank Hopfgartner, Benjamin Kille, Till Plumbaum, Andreas Lommatzsch, Torben Brodt, Arthur Bucko, and Tobias Heintz. Workshop and challenge on news recommender systems. In *RecSys'13: Proceedings of the International ACM Conference on Recommender Systems*. ACM, 10 2013.