GLARE 2018 – Generalization in Information Retrieval: Can We Predict Performance in New Domains?

Ian Soboroff
National Institute of Standards and Technology
Gaithersburg, Maryland, USA
ian.soboroff@nist.gov

Nicola Ferro
University of Padova
Padova, Italy
ferro@dei.unipd.it

Norbert Fuhr
University of Duisburg-Essen
Duisberg, Germany
norbert.fuhr@uni-due.de

1 MOTIVATION

A recent Dagstuhl Perspectives Workshop [2] tackled the problem of domain generalization: what are the barriers to being able to predict performance in new domains for NLP, IR, and recommender systems. This problem has been subsequently raised also during the SWIRL 2018 brainstorming workshop as one of the prominent issues in the next 5 years research agenda in IR [1]. Indeed, research in IR puts a strong focus on evaluation, with many past and ongoing evaluation campaigns. However, most evaluations utilize offline experiments with single queries only, while most IR applications are interactive, with multiple queries in a session. Moreover, context (e.g., time, location, access device, task) is rarely considered. Finally, the large variance of search topic difficulty make performance prediction especially hard.

Several types of prediction may be relevant in IR. One case is that we have a system and a collection and we would like to know what happens when we move to a new collection, keeping the same kind of task. In another case, we have a system, a collection, and a kind of task, and we move to a new kind of task. A further case is when collections are fluid, and the task must be supported over changing data.

Current approaches to evaluation mean that predictability can be poor, in particular:

- Assumptions or simplifications made for experimental purposes may be of unknown or unquantified validity; they may be implicit. Collection scale (in particular, numbers of queries) may be unrealistically small or fail to capture ordinary variability.
- Test collections tend to be specific, and to have assumed use-cases; they are rarely as heterogeneous as ordinary search. The processes by which they are constructed may rely on hidden assumptions or properties.
- Test environments rarely explore cases such as poorly specified queries, or the different uses of repeated queries (refinding versus showing new material versus query exploration, for example). Characteristics such as "the space of queries from which the test cases have been sampled" may be undefined.
- Researchers typically rely on point estimates for the performance measures, instead of giving confidence intervals. Thus, we are not even able to make a prediction about the results for another sample from the same population. A related confound is that highly correlated measures (for example, Mean Average Precision (MAP) vs normalized Discounted Cumulative Gain (nDCG)) are reported as if they were independent; while, on the other hand, measures which reflect different quality aspects (such as precision and recall) are averaged (usually with a harmonic mean), thus obscuring their explanatory power.
- Current analysis tools are focused on sensitivity (differences between systems) rather than reliability (consistency over queries).
- Summary statistics are used to demonstrate differences, but the differences remain unexplained. Averages are reported without analysis of changes in individual queries.

Perhaps the most significant issue is the gap between offline and online evaluation. Correlations between system performance, user behavior, and user satisfaction are not well understood, and offline predictions of changes in user satisfaction continue to be poor because the mapping from metrics to user perceptions and experiences is not well understood.

2 THEME

Following the manifesto drafted at Dagstuhl, we solicited research papers on the following topics:

(1) Measures: We need a better understanding of the assumptions and user perceptions underlying different metrics, as a basis for judging about the differences between methods. Especially, the current practice of concentrating on global measures should be replaced by using sets of more specialized metrics, each emphasizing certain perspectives or properties. Furthermore, the relationships between system-oriented and user-/task-oriented evaluation measures should be determined, in order to obtain a better improved prediction of user satisfaction and attainment of end-user goals.
(2) Performance analysis: Instead of regarding only overall performance figures, we should develop rigorous and systematic evaluation protocols focused on explaining performance differences. Failure and error analysis should aim at identifying general problems, avoiding idiosyncratic behavior associated with characteristics of systems or data under evaluation.
(3) Assumptions: The assumptions underlying our algorithms, evaluation methods, datasets, tasks, and measures should be identified and explicitly formulated. Furthermore, we need strategies for determining how much we are departing from them in new cases.
(4) Application features: The gap between test collections and real-world applications should be reduced. Most importantly,

we need to determine the features of datasets, systems, contexts, tasks that affect the performance of a system.

(5) Performance Models: We need to develop models of performance which describe how application features and assumptions affect the system performance in terms of the chosen measure, in order to leverage them for prediction of performance.

3 FORMAT

We received 11 submissions out of which 5 were accepted for final publication and presentation. The workshop itself will feature a keynote, an introductory talk presenting the Dagstuhl manifesto, research paper and position statement presentations. The final session of the day, in work groups, will prepare a research and data agenda to extend over the next several years.

Further information on the workshop are available on its Web site at: http://glare2018.dei.unipd.it/.

4 PROGRAM COMMITTEE

- Javed A. Aslam, Northeastern University, USA
- Ben Carterette, University of Delaware, USA
- Eric Gaussier, University Grenoble Alps, France
- Julio Gonzalo, UNED, Spain
- Gregory Grefenstette, INRIA Saclay – Ile-de-France, France
- Diane Kelly, University of Tennessee, USA
- Joseph A. Konstan, University of Minnesota, USA
- Claudio Lucchese, Ca' Foscari University of Venice, Italy
- Maria Maistro, University of Padua, Italy
- Josiane Mothe, University of Toulouse, France
- Jian-Yun Nie, Université de Montréal, Canada
- Raffaele Perego, ISTI CNR Pisa, Italy
- Gianmaria Silvello, University of Padua, Italy
- Ellen Voorhees, National Institute of Standards and Technology (NIST), USA
- Arjen P. de Vries, Radboud University, The Netherlands
- Justin Zobel, University of Melbourne, Australia

5 ACKNOWLEDGEMENTS

We would like to express our special thanks to the Program Committee members, the authors and all the attendees.

REFERENCES

[1] J. Allan, J. Arguello, L. Azzopardi, P. Bailey, T. Baldwin, K. Balog, H. Bast, N. Belkin, K. Berberich, B. von Billerbeck, J. Callan, R. Capra, M. Carman, B. Carterette, C. L. A. Clarke, K. Collins-Thompson, N. Craswell, W. B. Croft, J. S. Culpepper, J. Dalton, G. Demartini, F. Diaz, L. Dietz, S. Dumais, C. Eickhoff, N. Ferro, N. Fuhr, S. Geva, C. Hauff, D. Hawking, H. Joho, G. J. F. Jones, J. Kamps, N. Kando, D. Kelly, J. Kim, J. Kiseleva, Y. Liu, X. Lu, S. Mizzaro, A. Moffat, J.-Y. Nie, A. Olteanu, I. Ounis, F. Radlinski, M. de Rijke, M. Sanderson, F. Scholer, L. Sitbon, M. D. Smucker, I. Soboroff, D. Spina, T. Suel, J. Thom, P. Thomas, A. Trotman, E. M. Voorhees, A. P. de Vries, E. Yilmaz, and G. Zuccon. 2018. Research Frontiers in Information Retrieval – Report from the Third Strategic Workshop on Information Retrieval in Lorne (SWIRL 2018). *SIGIR Forum* 52, 1 (June 2018), 34–90.

[2] N. Ferro, N. Fuhr, G. Grefenstette, J. A. Konstan, P. Castells, E. M. Daly, T. Declerck, M. D. Ekstrand, W. Geyer, J. Gonzalo, T. Kuflik, K. Lindén, B. Magnini, J.-Y. Nie, R. Perego, B. Shapira, I. Soboroff, N. Tintarev, K. Verspoor, M. C. Willemsen, and J. Zobel. 2018. The Dagstuhl Perspectives Workshop on Performance Modeling and Prediction. *SIGIR Forum* 52, 1 (June 2018), 91–101.