Report on GLARE 2018: 1st Workshop on Generalization in Information Retrieval: Can We Predict Performance in New Domains?

Ian Soboroff
National Institute of Standards and Technology (NIST), USA
ian.soboroff@nist.gov

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

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

Abstract

This is a report on the first edition of the *International Workshop on Generalization in Information Retrieval* (GLARE 2018), co-located with the 27th ACM International Conference on Information and Knowledge Management (CIKM 2018) held in Turin, Italy, on October 22, 2018.

1 Motivations and Goals

GLARE 2018, the 1st International Workshop on Generalization in Information Retrieval was co-located with the 27th ACM International Conference on Information and Knowledge Management (CIKM 2018) held in Turin, Italy, on October 22, 2018.

A recent Dagstuhl Perspectives Workshop [4, 5] 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 (re-finding 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.

The workshop received 11 submissions out of which 5 were accepted for publication and presentation: 2 full papers and 3 position papers; Section 3 provides a short summary of the presented papers. All the presentations and papers discussed during the workshop are available on the GLARE 2018 Web site[1].

The workshop enjoyed an audience of more than 25 participants, who actively participated to the discussions fostered by the paper presentations and the keynote talk. The workshop also had a wide industrial participation with presenters and attendees from Amazon, Microsoft, Yahoo! Japan, and Yandex.

3 Paper Presentations

The Challenges of Moving from Web to Voice in Product Search

Ingber et al. [7] noted that voice has emerged as a disruptive user interface and is becoming pervasive in our homes, cars, and phones. One of the biggest challenges in this domain is searching and shopping by voice, which they refer to as "Voice Shopping". The voice shopping business is predicted to grow by at least an order of magnitude by 2020, from its $2 billion market today. Intuitively, one would think that it is possible to directly transfer the rich amount of data and signals from Web e-commerce domain to voice shopping. If feasible, this could enable a more efficient process of models development and validation. However, voice shopping, while directly related to Web e-commerce, introduces a new experience.

Ingber et al. [7] made the community aware that Web methods cannot be directly applied to the voice domain and they listed some research directions: (i) characterize and contrast users' behavior in Web vs voice product search; (ii) revisit user search experience in the voice domain; (iii) explore transfer learning methods from Web to voice.

[1] `http://glare2018.dei.unipd.it/`

Offline vs. Online Evaluation in Voice Product Search

Ingber et al. [8] observed that voice product search introduces a new paradigm and drives users' behavior to drastically differ from other domains with closed collections such as Web Mail or Web Product search. Before a ranking model is pushed into production, a common practice is first to evaluate it offline. However, Ingber et al. [8] argued that traditional search evaluation methods cannot be used "as is" in voice product search. They have shown that log-based offline experiments do not sufficiently correlate with online results to be valuable. Besides, voice shopping still being a new habit, online experiments might be riskier than in other environments such as Web search, as negatively affected users might not try the experience again.

Ingber et al. [8] encourage the community to explore new research directions for adequate training and evaluation of voice product search and they raised several challenges: (i) Can log-based data be leveraged when users are still learning a new medium and their behavior changes fast? (ii) Can manual golden sets enrich log-based data sets? (iii) Can log-based data be de-biased? (iv) Can we leverage data from random experiments?

Causality, prediction and improvements that (don't) add up

Fuhr [6] discussed how the development of causal models can be a promising approach to predicting performance of IR systems. Fuhr [6] noted how recent work on modelling performance of IR systems has developed models where the residual error is still large or where improvements did not add up linearly. He assumed that the major reason for this outcome is the problem that there are strong dependencies between the different methods, which were never investigated. Using such a causal model instead, it is possible to investigate the influence of the different variables on the final quality, and we are also able to regard the intermediate steps. Moreover, causal methods may be a helpful tool for performance prediction. Besides a clear notion of the variables affecting performance (and their interdependencies), we also need observable intermediate variables, that help us in the analysis

Towards a Basic Principle for Ranking Effectiveness Prediction without Human Assessments: A Preliminary Study

Amigó et al. [2] discussed a preliminary study about the Observational Information Linearity (OIL) assumption as a basic principle that explains the accuracy of pseudo relevance assessments in the IR literature. The proposed model predicts the effectiveness drop curve along positions in a single ranking, the relative performance of two rankings, and converges into the traditional pseudo assessment method when considering multiple rankings and statistical independence between them. Their proposed approach also allows for estimating the behavior of single rankings and the relative effectiveness of ranking pairs without having human assessments.

Novel Query Performance Predictors and their Correlations for Medical Applications

Bahrani and Roelleke [3] recognized that an obstacle to fully leverage Query Performance Prediction (QPP) is uncertainty in the effectiveness of the retrieval predictors when applied to different applications of the same task.They proposed novel pre-retrieval predictors that provide formal

grounds for the development of a probabilistic framework which serves QPP with respect to various IR models. They explored the influences of different representations of information needs on forecasting the retrieval quality concerning the medical collections. Their study discussed the role of Average Term Frequency, Inverse Document Frequency and the dependency between the query terms in the prediction. They used Dirichlet Multinomial and Natural Harmony assumption to develop new predictors which give rise to the term dependence assumption. Furthermore, they empowered the QPP tasks with a position-based TF-IDF measure which potentially enhances the prediction accuracy.

Acknowledgements

The workshop organizers are sincerely grateful to all the program committee members – 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).

Finally, we want to thank all the authors, speakers, and other participants for making GLARE 2018 a success.

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. Research Frontiers in Information Retrieval – Report from the Third Strategic Workshop on Information Retrieval in Lorne (SWIRL 2018). *SIGIR Forum*, 52(1):34–90, June 2018.

[2] E. Amigó, S. Mizzaro, and D. Spina. Towards a Basic Principle for Ranking Effectiveness Prediction without Human Assessments: A Preliminary Study. In I. Soboroff, N. Ferro, and N. Fuhr, editors, *Proc. 1st International Workshop on Generalization in Information Retrieval (GLARE 2018)*. `http://glare2018.dei.unipd.it/paper/glare2018-paper1.pdf`, 2018. Slides at: `http://glare2018.dei.unipd.it/talk/OIT_FOUNDATIONS_GLARE.pdf`.

[3] M. Bahrani and T. Roelleke. Novel Query Performance Predictors and their Correlations for Medical Applications. In I. Soboroff, N. Ferro, and N. Fuhr, editors, *Proc. 1st International Workshop on Generalization in Information Retrieval (GLARE 2018)*. `http://glare2018.dei.unipd.it/paper/glare2018-paper3.pdf`, 2018. Slides at `http://glare2018.dei.unipd.it/talk/GLARE_2018_Presentation.pdf`.

[4] 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. Manifesto from Dagstuhl Perspectives Workshop 17442 – From Evaluating to Forecasting Performance: How to Turn Information Retrieval, Natural Language Processing and Recommender Systems into Predictive Sciences. *Dagstuhl Manifestos, Schloss Dagstuhl–Leibniz-Zentrum für Informatik, Germany*, 7(1), 2018.

[5] 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. The Dagstuhl Perspectives Workshop on Performance Modeling and Prediction. *SIGIR Forum*, 52 (1):91–101, June 2018.

[6] N. Fuhr. Causality, prediction and improvements that (don't) add up. In I. Soboroff, N. Ferro, and N. Fuhr, editors, *Proc. 1st International Workshop on Generalization in Information Retrieval (GLARE 2018)*. `http://glare2018.dei.unipd.it/paper/glare2018-paper2.pdf`, 2018. Slides at `http://glare2018.dei.unipd.it/talk/glare18-talk-fuhr.pdf`.

[7] A. Ingber, A. Lazerson, L. Lewin-Eytan, A. Libov, and E. Osherovich. The Challenges of Moving from Web to Voice in Product Search. In I. Soboroff, N. Ferro, and N. Fuhr, editors, *Proc. 1st International Workshop on Generalization in Information Retrieval (GLARE 2018)*. `http://glare2018.dei.unipd.it/paper/glare2018-paper5.pdf`, 2018. Slides at `http://glare2018.dei.unipd.it/talk/CIKM-GLARE-VoiceVsWeb_v2.pdf`.

[8] A. Ingber, L. Lewin-Eytan, A. Libov, Y. Maarek, and E. Osherovich. Offline vs. Online Evaluation in Voice Product Search. In I. Soboroff, N. Ferro, and N. Fuhr, editors, *Proc. 1st International Workshop on Generalization in Information Retrieval (GLARE 2018)*. `http://glare2018.dei.unipd.it/paper/glare2018-paper4.pdf`, 2018. Slides at `http://glare2018.dei.unipd.it/talk/CIKM-GLARE-OnlineVsOffline_v2.pdf`.